

# Complete intersection for equivariant models

M. Casanellas <sup>\*1</sup>, J. Fernández-Sánchez <sup>†1</sup>, and M. Michałek <sup>‡2</sup>

<sup>1</sup>Departament de Matemàtiques, Universitat Politècnica de Catalunya

<sup>2</sup>Freie Universität, Berlin, Germany; UC Berkeley, USA; Polish Academy of Sciences, Warsaw

December 23, 2015

## Abstract

Phylogenetic varieties related to equivariant substitution models have been studied largely in the last years. One of the main objectives has been finding a set of generators of the ideal of these varieties, but this has not yet been achieved in some cases (for example, for the general Markov model this involves the open “salmon conjecture”, see [2]) and it is not clear how to use all generators in practice. Motivated by applications in biology, we tackle the problem from another point of view. The elements of the ideal that could be useful for applications in phylogenetics only need to describe the variety around certain *points of no evolution* (see [13]). We produce a collection of explicit equations that describe the variety on a Zariski open neighborhood of these points (see Theorem 5.4). Namely, for any tree  $T$  on any number of leaves (and any degrees at the interior nodes) and for any equivariant model on any set of states  $\kappa$ , we compute the codimension of the corresponding phylogenetic variety. We prove that this variety is smooth at general points of no evolution, and provide an algorithm to produce a complete intersection that describes the variety around these points.

## 1 Introduction

In the recent years there has been a huge amount of work done on phylogenetic varieties – we advise the reader to consult e.g. [4, 7, 12, 27, 39, 44] and references therein. These algebraic varieties contain the set of joint distributions at the leaves of a tree evolving under a Markov model of molecular evolution. From the biological point of view, these varieties are interesting because they provide new tools of non-parametric inference of phylogenetic trees. At present, the algebraic/geometric framework of phylogenetic varieties has allowed proving the identifiability of parameters of certain evolutionary models widely used by biologists [8, 3], proposing new methods of model selection [34], and producing new phylogenetic reconstruction methods [28, 15].

From the mathematical point of view, there has been a great effort in finding a whole description of the ideal of these phylogenetic varieties [6, 44, 24, 23]. Still, for some models, many questions remain open for trees on an arbitrary number of leaves  $n$ . Indeed, if one is interested in using these algebraic tools with real data, one would need a small set of generators of the ideal (rather than a description of the whole ideal); it would also be desirable to know the degree at which

---

<sup>\*</sup>marta.casanellas@upc.edu

<sup>†</sup>jesus.fernandez.sanchez@upc.edu

<sup>‡</sup>wajcha@berkeley.edu

the ideal is generated [25, 40, 44]; and as the codimension of the variety is exponential in  $n$ , it is necessary to distinguish between generators that only account for the underlying evolutionary model, those that account for the tree topology, and those that could be useful for inferring the numerical parameters (see [7] for a good introduction to this topic).

For instance, the authors of [44] and [7] raised the question whether knowing complete intersection containing the phylogenetic variety and of the same dimension would be enough. Eminently, for biological applications it is only relevant to know the description of the variety around the points that make sense *biologically* speaking. If these points are smooth, then a complete intersection can define the variety on a neighborhood of these points.

This is the approach that was considered in [13, 32, 41] for the particular model of Kimura 3-parameter and is the same goal that we pursued in [18] for abelian group-based models. In the present paper we address this problem for the more general class of *G-equivariant models*, which contains more general algebraic models of interest to biologists (for example the strand-symmetric and the general Markov models). We give an explicit algorithm to construct a complete intersection that describes the variety on a dense open subset around a generic *point of no evolution*. Points of no evolution represent molecular sequences that remain invariant from the common ancestor to the leaves of the tree. Points in the phylogenetic variety that arise from biologically meaningful parameters are supposed to be near these points of no evolution (otherwise phylogenetic inference could not be made), and therefore it is important to study the variety around these points. Also, as we describe the variety at a dense open subset containing these points, we cover most (actually, all but possibly a subset of smaller dimension) of the biologically meaningful points of the variety. Also, in the same papers mentioned above, it is argued that a complete intersection can contain points in other irreducible components that can mislead the results in practice. However, the complete intersection we give contains a regular sequence of the edge invariants, which are known to be phylogenetic invariants [14]. Therefore, the other irreducible components of the complete intersection do not contain other phylogenetic varieties.

We prove first that these points are non-singular and therefore the variety can be described locally at these points by the smallest possible number of equations, the codimension of the variety. A system of generators of the local complete intersection can be explicitly computed. The degree of these generators is low and depends on a local description of claw trees related to the interior nodes of the tree and of the multiplicities of the permutation representation of the group  $G \subset \mathfrak{S}_\kappa$ . For example, for the biologically interesting models mentioned above, the complete intersection we provide has generators of degree at most 13. One should contrast this to the generators of the complete intersection given in [43] for the Kimura 3-parameter model, which had exponential degree in the number of leaves. Our approach is also useful in case one wants to use differential geometry for this variety (for example to compute the distance of a point to this variety, [26]).

The description of the ideal of phylogenetic invariants for *G-equivariant models* was provided by Draisma and Kuttler [24]. There are two ways to obtain the whole ideal of phylogenetic invariants for a given tree, assuming the ideals for star trees are known. The first description relies on the ideals of star trees associated to inner vertices of a given tree - so-called flattenings. The second description is inductive, where we regard a big tree as a join of two smaller trees.

Our approach is based on the second method. We start by inducing phylogenetic invariants from smaller trees. The induced phylogenetic invariants are of course not enough to provide a description for the larger tree. We complement them by so-called thin flattenings [14]. They are very explicit, however still numerous. It turns out that the choice of leaves in smaller trees distinguishes specific thin flattenings. Combining those with induced invariants yields our main result: under a minor assumption on claw trees (see 5.2) which is satisfied by the tripod on the most popular equivariant models (Jukes-Cantor, Kimura 3ST, strand symmetric) and also by the general Markov model, we provide an explicit local description of the variety associated to a model and a tree (see Theorem 5.4). Moreover, both in the starting point and in the induction process,

the choices we make are almost canonical so that the complete intersection we produce is a natural one and could be reasonably used in practice.

The methods used in this paper rely on basic algebraic geometry and group representation theory. It is important to note that the results of the paper hold for any  $G$ -equivariant model,  $G \subset \mathfrak{S}_\kappa$ , for any  $\kappa$ , and therefore representation theory has been the necessary tool to deal with all these models at the same time. On the other hand, our results also hold for trees with any number of leaves and any degrees at the interior nodes.

The approach adopted to prove the main result 5.4 also produces a computationally effective list of elements in the ideal of the phylogenetic variety. Indeed, the list of equations provided in Theorem 5.4 for a tree  $T$  with  $n$  leaves is constructed from equations describing locally the phylogenetic varieties of claw trees of the interior nodes of  $T$  and from certain minors of the thin flattenings mentioned above. The number of equations from the thin flattenings grows exponentially with  $n$ , but the number of equations corresponding to claw trees does not (it grows exponentially with the maximum degree of an interior node of  $T$ , see Remarks 4.7 and 5.6). However, evaluating the minors of the thin flattenings is not the optimal way of evaluating the rank of a matrix and these equations could therefore be substituted by a numerical method such as the singular value decomposition (see [28]). The remaining equations form a set that can be useful in practice, for example for the estimation of the parameters that maximize the likelihood via Lagrange multipliers (see the tools used in [22] and [21]).

The structure of the paper is as follows. In the following section 2, we recall the background on linear representation theory of finite groups that is needed in the sequel. In section 3, we recall the definition of equivariant models and of phylogenetic varieties. In this section we prove as well two key results that shall allow us to provide a complete intersection as desired: first we compute the dimension of the phylogenetic varieties for any equivariant model  $\mathcal{M}_G$ ,  $G \subset \mathfrak{S}_\kappa$ , and any tree  $T$ , and then we prove that these varieties are smooth at generic points of no evolution. Then in section 4 we describe the set of equations that shall be used to prove our main result. The setup for this description is conceived towards the induction steps that are needed in the proof of the main result. In section 5 we describe the induction step and the “claw tree hypotheses” needed to prove our main result, Theorem 5.4 in the largest generality. The proof of this theorem is constructive and provides an algorithm for obtaining the desired complete intersection assuming the claw tree hypotheses is satisfied. In section 6 we prove that this claw tree hypothesis holds for trivalent trees on the general Markov model, the strand symmetric model, and the Jukes-Cantor model (the Kimura 3-parameter case was already considered in [13]). For these models we also specify complete intersections (following the algorithm provided in section 5) that describe the variety for quartet trees around generic points of no evolution.

## Acknowledgements

M.Casanellas and J.Fernández-Sánchez were partially supported by Spanish government MTM2012-38122-C03-01/FEDER and Generalitat de Catalunya 2014 SGR-634. M.Michalek was supported by National Science Center grant SONATA UMO-2012/05/D/ST1 /01063. Part of the work was conducted while Michalek was visiting Freie Universität, Berlin and UC Berkeley (PRIME DAAD program 2015-2016) and CRM Barcelona (EPDI program 2013).

## 2 Background on representation theory

In this section we recall the basic concepts of representation theory that will be needed in the sequel. The reader is referred to [42] or [31] for details and proofs. Throughout the paper we work

over the field of complex numbers  $\mathbb{C}$ .

Let  $G$  be a finite group. A *representation* of  $G$  is a group homomorphism  $\rho : G \rightarrow GL(V)$ , where  $V$  is a  $\mathbb{C}$ -vector space of finite dimension. We will refer to  $V$  as the representation itself (or also as a  $G$ -module) if the map  $\rho$  can be understood from the context, and for  $g \in G$  and  $u \in V$  we shall denote by  $gu$  the vector  $\rho(g)(u)$ . A  $G$ -*equivariant map* is a linear map  $f : V \rightarrow V'$  between two representations of  $G$  that satisfies  $f \circ \rho(g) = \rho_{V'}(g) \circ f$  for all  $g \in G$ . The set of all  $G$ -equivariant maps between  $V$  and  $V'$  is denoted as  $\text{Hom}_G(V, V')$ . Two  $G$ -modules  $V$  and  $V'$  are said to be *isomorphic* (denoted as  $V \cong_G V'$ ) if there is a  $G$ -equivariant isomorphism of vector spaces  $f : V \rightarrow V'$ . A representation  $V$  is *irreducible* if it does not contain any proper  $G$ -invariant subspace. Otherwise,  $V$  is said to be *reducible*. We will denote by  $V^G$  the subspace of vectors of  $V$  that are  $G$ -invariant under the action of  $G$ , that is,  $gu = u$  for all  $g \in G$ .

**Lemma 2.1 (Schur)** *Let  $V, V'$  be two irreducible representations of  $G$ . If  $f : V \rightarrow V'$  is  $G$ -equivariant, then either  $f = 0$  or  $f$  is an isomorphism, in which case  $\text{Hom}_G(V, V') \cong \mathbb{C}$ .*

**Notation 2.2** Let  $\rho_k : G \rightarrow GL(N_k)$ ,  $k = 1, \dots, t$  be the irreducible representations of  $G$  (up to isomorphism). We write  $\chi_k$  for the character corresponding to  $N_k$ :  $\chi_k(\sigma) = \text{trace}(\rho_k(\sigma))$ . We adopt the convention that  $(\rho_1, N_1)$  refers to the *identity* (or *trivial*) representation.

**Theorem 2.3 (Maschke)** *If  $\rho : G \rightarrow GL(V)$  is a representation of  $G$ , then there exists a unique decomposition  $V = \bigoplus_{k=1}^t V[\chi_k]$ , where each  $V[\chi_k]$  is isomorphic to  $\bigoplus^{m_k(V)} N_k$  for some multiplicity  $m_k(V) \geq 0$ . We call  $V[\chi_k]$  the isotypic component of  $V$  associated to  $N_k$ .*

Notice that  $V^G$  is equal to the isotypic component of  $V$  associated to the trivial representation of  $G$ :  $V^G = V[\chi_1]$ .

**Remark 2.4** By virtue of these fundamental results, for any representation  $V$  of  $G$ , the dimension of  $\text{Hom}_G(N_k, V)$  equals the multiplicity of  $N_k$  in  $V$ . Moreover, the collection of the images of a chosen vector  $v_k \in N_k$  under maps in  $\text{Hom}_G(N_k, V)$  form a subspace  $\mathcal{F}_k(V)$  in  $V$  of dimension equal to the multiplicity of the isotypic component,  $\mathcal{F}_k(V) \cong \mathbb{C}^{m_k(V)}$ . Analogously to highest weight spaces, the spaces  $\mathcal{F}_k(V)$  will represent the whole isotypic components. In particular, for any two representations  $V, V'$  we can identify  $\text{Hom}_G(V, V')$  with  $\bigoplus_k \text{Hom}_{\mathbb{C}}(\mathcal{F}_k(V), \mathcal{F}_k(V'))$ .

**Permutation representation.** From now on we focus on the following setting. Given a finite set  $\Sigma$  of cardinality  $\kappa$ , we define  $W = \langle \Sigma \rangle_{\mathbb{C}}$  as the  $\mathbb{C}$ -vector space generated by the elements of  $\Sigma$ . In this way, the elements of  $\Sigma$  play the role of the standard basis of  $W$ , so that an element  $\mathbf{x} \in \Sigma$  and the corresponding vector of the standard basis shall be denoted in the same way. Motivated by biology, in our examples we consider  $\Sigma = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$  but our work holds for any finite set. We denote  $\mathbf{1} := \sum_{\mathbf{x} \in \Sigma} \mathbf{x}$ . By abuse of notation,  $\mathbf{1}$  will be sometimes taken as the column-vector with all its  $\kappa$  coordinates equal to one. Henceforth,  $G$  shall be a permutation group of  $\Sigma$ , that is,  $G$  is a subgroup of  $\mathfrak{S}_{\kappa}$ . The restriction to  $G$  of the *permutation representation*  $W$ , given by the permutation of the elements in  $\Sigma$ , induces a representation  $\rho(s)$  of  $G$  on any tensor power  $\otimes^s W$  by extending linearly the action  $\sigma(\mathbf{x}_{i_1} \otimes \dots \otimes \mathbf{x}_{i_s}) := \sigma \mathbf{x}_{i_1} \otimes \dots \otimes \sigma \mathbf{x}_{i_s}$  for  $\sigma \in G, \mathbf{x}_{i_j} \in \Sigma$ . In this paper, we will only deal with such representations  $\rho(s) : G \rightarrow GL(\otimes^s W)$  together with the irreducible representations  $N_1, \dots, N_t$  of  $G$ .

According to Maschke's theorem, any tensor power  $\otimes^s W$  will decompose into a direct sum of modules  $(\otimes^s W)[\chi_k]$  (the isotypic components) each of them being a number of copies of one of the irreducible modules  $N_k$ . This number is the multiplicity of  $N_k$  in  $\otimes^s W$  and will be denoted by  $m_k(s)$ . In the particular case  $G = \mathfrak{S}_k$ , explicit formulas for  $m_k(s)$  can be provided in terms of Kronecker coefficients. We write  $\mathbf{m}(s) = (m_1(s), \dots, m_t(s))$  for the vector of multiplicities of

$\otimes^s W$ . As the case  $s = 1$  will play a special role, we simplify notation and write  $\mathbf{m} = (m_1, \dots, m_t)$  for the vector of multiplicities of  $W$ .

From now on, we fix subspaces  $\mathcal{F}_k(W) \subset W$  for  $k = 1, \dots, t$  according to Remark 2.4 by fixing a vector  $v_k \in N_k$  and taking its images by maps in  $\text{Hom}_G(N_k, W)$ . This vector also defines subspaces  $\mathcal{F}_k(\otimes^s W) \subset \otimes^s W$ , which shall be considered fixed from now on.

We consider the Hermitian inner product in  $W$  that makes  $\Sigma$  into an orthonormal basis, and denote it by  $v \cdot w$  for any  $v, w \in W$ . This inner product will be used to identify  $W$  with  $W^*$  by sending a vector  $v$  to the linear form  $v^* \in W^*$  that maps  $u$  to  $v \cdot u$ .

The inner product in  $W$  induces an inner product in  $\otimes^s W$  defined as

$$\mathbf{X}_1 \otimes \dots \otimes \mathbf{X}_n \cdot \mathbf{Y}_1 \otimes \dots \otimes \mathbf{Y}_n := \prod_{k=1}^s \mathbf{X}_k \cdot \mathbf{Y}_k,$$

for  $\mathbf{X}_i, \mathbf{Y}_j \in \Sigma$  and extending it sesquilinearly.

**Dual representation.** If  $\rho : G \longrightarrow GL(V)$  is a representation of  $G$  with character  $\chi$ , then its dual  $V^*$  is also a representation via the homomorphism  $\rho^* : G \longrightarrow GL(V^*)$  that maps  $g$  to  ${}^t\rho(g^{-1})$ , and  $V^*$  has character  $\chi^*$  (the conjugate of  $\chi$ ).

At the level of vector spaces, the inner product above provides an isomorphism  $V \cong V^*$ . Nevertheless, if  $V$  is a representation of a group  $G$ , then it may happen that it is not  $G$ -isomorphic to  $V^*$ . This will force us to distinguish between the space and its dual in the sequel. However, the permutation representations  $V = \otimes^s W$  that we consider in this paper satisfy  $V \cong_G V^*$  because they have real characters. In particular, the  $G$ -isomorphism  $V \cong_G V^*$  induces  $G$ -isomorphisms  $V^*[\chi_k] \cong_G V[\chi_k]$  and  $\mathcal{F}_k(V) \cong \mathcal{F}_k(V^*)$  for all  $k$ .

Thus, the reader may freely ignore all the dual signs in our article. We decided to keep them, as most of the arguments we provide hold without the assumption  $V \cong_G V^*$  on a representation theoretic level. There is also a natural  $G$ -isomorphism  $V \otimes V' \cong_G \text{Hom}(V^*, V')$  which at the level of  $G$ -invariant vectors translates to  $(V \otimes V')^G \cong_G \text{Hom}_G(V^*, V')$ .

Representation theory allows us to decompose the ambient space  $(\otimes^n W)^G$  in terms of the irreducible representations of  $G$  as follows. This decomposition will be fundamental for us and will play a key role in the paper.

**Proposition 2.5** *For any  $a + b = n$ , there is a natural isomorphism of vector spaces*

$$(\otimes^{a+b} W)^G \cong \bigoplus_{k=1}^t \mathcal{F}_{k^*}(\otimes^a W) \otimes \mathcal{F}_k(\otimes^b W).$$

*In particular, the dimension of  $(\otimes^{a+b} W)^G$  is  $m_1(a+b) = \sum_{k=1}^t m_{k^*}(a)m_k(b)$ , where  $k^*$  is the index of the irreducible representation dual to  $N_k$ , that is,  $N_{k^*} = (N_k)^*$ . Using the language of category theory, the functors  $(\otimes^n \cdot)^G$  and  $\bigotimes_{k=1}^t \mathcal{F}_{k^*}(\otimes^a \cdot) \otimes \mathcal{F}_k(\otimes^b \cdot)$  from the category of  $G$  representations to the category of vector spaces are isomorphic.*

PROOF. Applying Maschke's theorem and Schur's lemma, we infer

$$\begin{aligned} (\otimes^n W)^G &\cong ((\otimes^a W) \otimes (\otimes^b W))^G \cong \text{Hom}_G((\otimes^a W)^*, \otimes^b W) \\ &\cong \bigoplus_{i,j} \text{Hom}_G((\otimes^a W)^*[\chi_i], (\otimes^b W)[\chi_j]) \cong \bigoplus_{k=1}^t \text{Hom}_{\mathbb{C}}(\mathcal{F}_k(\otimes^a W^*), \mathcal{F}_k(\otimes^b W)) \\ &\cong \bigoplus_{k=1}^t \text{Hom}_{\mathbb{C}}((\mathcal{F}_{k^*}(\otimes^a W))^*, \mathcal{F}_k(\otimes^b W)) \cong \bigoplus_{k=1}^t \mathcal{F}_{k^*}(\otimes^a W) \otimes \mathcal{F}_k(\otimes^b W). \end{aligned}$$

□

### 3 Equivariant evolutionary models and phylogenetic varieties

A tree is a connected finite graph without cycles, consisting of vertices and edges. Given a tree  $T$ , we write  $V(T)$  and  $E(T)$  for the set of vertices and edges of  $T$ . The *degree* of a vertex is the number of edges incident to it. The set  $V(T)$  splits into the set of leaves  $L(T)$  (vertices of degree one) and the set of interior vertices  $Int(T)$  :  $V(T) = L(T) \cup Int(T)$ . One says that a tree is *trivalent* if each vertex in  $Int(T)$  has degree 3. A tree topology is the topological class of a tree where every leaf has been labeled. Given a subset  $A$  of  $L(T)$ , the subtree induced by  $A$  is just the smallest tree composed of the edges and vertices of  $T$  in any path connecting two leaves in  $A$ . A tree  $T$  is *rooted* if a specific node  $r$  is labeled as the root.

In order to model the substitution of the states in  $\Sigma$  according to a Markov process on a rooted tree  $T$ , one has to specify a distribution  $\pi$  at the root of the tree and a collection of substitution matrices  $\mathbf{A} = (A^e)_{e \in E(T)}$  [19, 16]. The set of possible root distributions and substitution matrices for a tree  $T$  is called the *set of parameters*. In the applications to biology, one has to restrict the set of parameters to stochastic vectors and matrices, but this restriction is unnecessary for the core of this paper. Below we describe *equivariant* models of evolution, which include some of the most well-known models.

As above, let  $\Sigma$  be a finite set of cardinal  $\kappa$ ,  $W$  be the  $\mathbb{C}$ -vector space  $\langle \Sigma \rangle_{\mathbb{C}}$ , and  $G \leq \mathfrak{S}_{\kappa}$  be a permutation group of  $\Sigma$ . In this section we use the distinguished basis  $\Sigma$  of  $W$  to identify  $\kappa \times \kappa$  matrices with complex entries with  $\text{Hom}(W, W)$ .

**Definition 3.1** (cf. [24]) A rooted tree  $T$  evolves under the *equivariant model*  $\mathcal{M}_G$  if a  $G$ -invariant vector  $\pi$  is associated to the root of  $T$  and substitution matrices  $A^e$  in  $\text{Hom}_G(W, W)$  are associated to each edge  $e$  of  $T$ . For the equivariant model  $\mathcal{M}_G$ , the set of *parameters* is

$$\text{Par}_G(T) = W^G \times \prod_{e \in E(T)} \text{Hom}_G(W, W).$$

If one wants to talk about *stochastic parameters*,  $s\text{Par}_G(T)$  one has to restrict the root distribution to  $sW^G := W^G \cap \{\pi \in W \mid \pi \cdot \mathbf{1} = 1\}$ , and the substitution matrices to  $s\text{Hom}_G(W, W) := \text{Hom}_G(W, W) \cap \{A \mid A \cdot \mathbf{1} = \mathbf{1}\}$  (and then require that all entries are real, nonnegative, but this is not relevant for our purposes). As a special case, if the group  $G$  is trivial,  $G = \{1\}$ , we will denote by  $\text{Par}(T)$  and  $s\text{Par}(T)$  the corresponding spaces of parameters. The *parametrization map* that assigns a distribution at the leaves of  $T$  to each set of parameters is

$$\Psi_T : \text{Par}(T) \longrightarrow \otimes^n W \tag{1}$$

defined by

$$\Psi_T(\pi, \mathbf{A}) = \sum_{\mathbf{x}_i \in \Sigma} p_{\mathbf{x}_1 \dots \mathbf{x}_n} \mathbf{x}_1 \otimes \dots \otimes \mathbf{x}_n,$$

where

$$p_{\mathbf{x}_1 \dots \mathbf{x}_n} = \sum_{\mathbf{x}_v \in \Sigma, v \in Int(T)} \pi_{\mathbf{x}_r} \prod_{e \in E(T)} A_{\mathbf{x}_{pa(e)}, \mathbf{x}_{ch(e)}}^e, \tag{2}$$

$\mathbf{x}_v$  denotes the state at the vertex  $v$ ,  $pa(e)$  (respectively  $ch(e)$ ) is the parent (respectively, child) node of  $e$ , and  $(\pi_{\mathbf{x}})_{\mathbf{x} \in \Sigma}$  are the coordinates of the root distribution  $\pi$ . When we restrict this map to the set of parameters  $\text{Par}_G(T)$ , we denote it as  $\Psi_T^G$ . In this case the image lies in  $(\otimes^n W)^G$ .

When the parametrization (1) is restricted to the set of stochastic parameters, we obtain

$$\phi_T : s\text{Par}(T) \longrightarrow H \cap \otimes^n W,$$

where  $H \subset \otimes^n W$  is the hyperplane defined as

$$H = \left\{ p \in \otimes^n W \mid \sum_{\mathbf{x}_i \in \Sigma} p_{\mathbf{x}_1 \dots \mathbf{x}_n} = 1 \right\}.$$

The analogous restrictions to  $s\text{Par}_G(T)$  are denoted as  $\phi_T^G$ . The word “stochastic” here has a broader meaning than usually, because for our aim we only need entries summing to one and not necessarily nonnegative entries.

The *phylogenetic variety associated to a tree  $T$  evolving under  $\mathcal{M}_G$*  is the (affine) algebraic variety

$$CV_G(T) := \overline{\{\Psi_T^G(\pi, \mathbf{A}) \mid (\pi, \mathbf{A}) \in \text{Par}_G(T)\}} \subset (\otimes^n W)^G.$$

where  $\overline{S}$  represents the Zariski closure of a set  $S$ . Similarly, the *stochastic phylogenetic variety associated to a tree  $T$*  is the smallest algebraic variety  $V_G(T)$  containing the set

$$\text{Im } \phi_T^G = \{\phi_T^G(\pi, \mathbf{A}) : (\pi, \mathbf{A}) \in s\text{Par}_G(T)\}.$$

One has  $V_G(T) = CV_G(T) \cap H$  (see, for example, [16]). In particular, the equations defining  $V_G(T)$  are the same equations defining  $CV_G(T)$  plus the equation defining  $H$ .

Notice that in the definition of the phylogenetic variety we have not specified the root of the tree. It is well known that different root placements give rise to the same phylogenetic variety. Indeed, it is clear that a matrix  $M$  belongs to  $\text{Hom}_G(W, W)$  if and only if so does its transpose  $M^t$ . Now, it can be seen that if we move the root from one node to a neighboring node and we replace the matrices  $A^e$  of the edges with inverted orientation with their transpose, the image of the new parameters will remain the same. Moreover, we may assume that  $\pi = \mathbf{1}$  when parameterizing  $CV_G(T)$  since choosing an edge  $e_0$  attached to the root and changing  $A^{e_0}$  by  $\text{diag}(\pi)A^{e_0}$  gives rise to the same image point. Hence  $CV_G(T)$  does not depend on the root. As  $H$  also does not depend on the root, neither does  $V_G(T)$ .

In case we take all matrices  $A^e$  equal to the identity, the image by  $\Psi_T^G$  represents no evolution at all.

**Definition 3.2** Given an equivariant model  $\mathcal{M}_G$ , a point  $\pi_n$  in  $\otimes^n W$  is a *point of no evolution* if  $\pi_n = \sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} \mathbf{x} \otimes \dots \otimes \mathbf{x}$  and  $\pi_n$  is invariant under  $G$ .

If  $\pi_n = \sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} \mathbf{x} \otimes \dots \otimes \mathbf{x}$  is a point of no evolution, then it belongs to  $CV_G(T)$  for any tree  $T$  on  $n$  leaves because  $\pi_n = \Psi_T^G(\pi, \mathbf{I})$  where  $\pi = \sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} \mathbf{x} \in W^G$  and  $\mathbf{I}$  corresponds to the identity matrix at each edge:  $\mathbf{I} = (Id)_{e \in E(T)}$ .

**Remark 3.3** For biological applications, we are interested in real/stochastic points in  $\otimes^n W$  that are close to points of no evolution (in the complex euclidean distance). Indeed, as  $\Psi_T^G$  is a continuous map, if  $p$  is close to a point of no evolution  $\pi_n$ , then there is a preimage of  $p$  close to  $(\pi, \mathbf{I})$ . The parameters close to  $(\pi, \mathbf{I})$  are precisely those that are interesting biologically speaking because they account for probabilities of no mutation greater than probabilities of mutation (that is, diagonal entries greater than off-diagonal entries in transition matrices). To this end, the main goal of this paper is to provide a local description of the phylogenetic varieties around the points of no evolution.

**Example 3.4** The definition of equivariant model includes important evolutionary models used in phylogenetics for  $\kappa = 4$  like

1. Jukes-Cantor [33], for  $G = \mathfrak{S}_4$  or  $G = \mathfrak{A}_4$  (the alternating group);
2. Kimura 2-parameter [35], for  $G = \langle (\text{ACGT}), (\text{AG}) \rangle$ ;
3. Kimura 3-parameter [36], for  $G = \langle (\text{AC})(\text{GT}), (\text{AG})(\text{CT}) \rangle$ ;
4. strand-symmetric [17], for  $G = \langle (\text{AT})(\text{CG}) \rangle$ ;
5. general Markov (briefly GMM) [10], for  $G = \{1\}$ .

We say that  $\mathcal{M}_G$  is a *submodel* of  $\mathcal{M}_H$  if  $H \leq G$ . With this terminology, all the models above are submodels of the general Markov model and we have inclusions from top to bottom on the sets of corresponding parameters (and algebraic varieties).

### 3.1 Dimension of phylogenetic varieties for equivariant models

In this subsection we compute the dimension of the phylogenetic variety associated to any  $G$ -equivariant model on any tree  $T$ ,  $G \leq \mathfrak{S}_\kappa$ . This dimension was already known in the particular cases of the Jukes-Cantor, Kimura 2 and 3 parameters and general Markov model. The result yields the codimension of these varieties and hence it is the first step towards providing a complete intersection containing them.

**Theorem 3.5** *For any group  $G \leq \mathfrak{S}_\kappa$  and any tree  $T$  without nodes of degree 2, the dimension of  $V_G(T)$  is  $|E(T)|(m_1(2) - m_1) + m_1 - 1$  and the dimension of  $CV_G(T)$  is  $|E(T)|(m_1(2) - m_1) + m_1$ .*

PROOF. The dimension of  $V_G(T)$  is upper bounded by the dimension of  $s\text{Par}_G(T)$ , which we compute in the following.

Each transition matrix  $M$  is an element of

$$\text{Hom}_G(W, W) \cong (W^* \otimes W)^G,$$

so, as in our case  $W \cong W^*$ , the number of parameters is  $m_1(2)$ . However, because of the stochastic assumption, the sum of the rows of each matrix  $M$  is fixed to one. Notice that  $(W^* \otimes W)^G$  surjects onto  $W^G$  by the map  $M \mapsto M\mathbf{1}$ . Hence, there are  $m_1$  independent restrictions on the parameters of  $M$ . This makes  $|E(T)|(m_1(2) - m_1)$  free parameters for the choice of the transition matrices. On the other hand, the distribution of the root is given by a vector  $\pi \in W^G$ . The stochastic condition implies that the sum of the coordinates is equal to one. This makes  $m_1 - 1$  free parameters for the choice of the root distribution.

Summing up, we have that  $\dim V_G(T)$  is less or equal than  $|E(T)|(m_1(2) - m_1) + (m_1 - 1)$ .

In order to prove the other inequality we use Chang's result ([19]) and its generalization ([4]) on the "generic identifiability of parameters" of the general Markov model  $\mathcal{M}_1$  on trees without nodes of degree 2. This result says that the fiber of  $\phi_T(\mathcal{P})$  is finite for parameters  $\mathcal{P} = (\pi, (A^e)_e)$  that satisfy: (1) no entry of  $\pi$  is zero; (2) all  $A^e$  are non-singular; and (3)  $\det A^e \neq \pm 1$  for all  $e$ . These generic conditions (1)-(3) are also generic for the parameters of any equivariant model  $\mathcal{M}_G$ . That is, if  $\Sigma = \{\mathbf{x}_1, \dots, \mathbf{x}_\kappa\}$ , for any group  $G \leq \mathfrak{S}_\kappa$  we have

- (i)  $sW^G$  is not included in  $\{\pi \in W \mid \pi_{\mathbf{x}_1} \cdot \dots \cdot \pi_{\mathbf{x}_\kappa} = 0\}$  (indeed,  $\frac{1}{\kappa}\mathbf{1} \in sW^G$  for example),
- (ii)  $s\text{Hom}_G(W, W)$  is not contained in the set of singular matrices (indeed,  $Id \in s\text{Hom}_G(W, W)$ ),  
and



- (iii)  $s\text{Hom}_G(W, W)$  does not only contain matrices with determinant 1 or  $-1$  (for example, the matrix with all entries equal to  $1/\kappa$  belongs to  $s\text{Hom}_G(W, W)$ ).

This means that for generic parameters  $\mathcal{P} \in s\text{Par}_G(T)$ , if we set  $p = \phi_T^G(\mathcal{P})$ , the preimage  $(\phi_T)^{-1}(p)$  is finite. As the preimage  $(\phi_T)^{-1}(p)$  contains  $(\phi_T^G)^{-1}(p)$ , this implies that the generic fiber of  $\phi_T^G$  is finite. Therefore, the dimension of  $V_G(T)$  is upper bounded by the dimension of the domain of  $\phi_T^G$ , which has been computed above.

The claim for  $CV_G(T)$  follows because it is the closure of the cone over  $V_G(T)$ .  $\square$

This result implies the generic identifiability of the stochastic parameters for trees evolving under equivariant models (see [9, Definition 1] for example).

**Corollary 3.6** *The stochastic parameters of a tree  $T$  evolving under an equivariant model  $\mathcal{M}_G$ ,  $G \leq \mathfrak{S}_\kappa$ , are generically identifiable if  $T$  has no nodes of degree 2.*

**Remark 3.7** It can be checked easily that for the evolutionary models listed in Example 3.4, the dimension for a trivalent tree on  $n$  leaves (and hence with  $2n - 3$  edges)  $T$  is

1.  $\dim_{\mathbb{C}} V_G(T) = 2n - 3$  for the Jukes-Cantor model;
2.  $\dim_{\mathbb{C}} V_G(T) = 4n - 6$  for the Kimura 2-parameter model;
3.  $\dim_{\mathbb{C}} V_G(T) = 6n - 9$  for the Kimura 3-parameter model;
4.  $\dim_{\mathbb{C}} V_G(T) = 12n - 17$  for the strand symmetric model;
5.  $\dim_{\mathbb{C}} V_G(T) = 24n - 33$  for the general Markov model.

If  $T$  has  $n$  leaves, we will write  $\text{codim}(T)$  for the codimension of  $CV_G(T)$  in  $(\otimes^n W)^G$  (equal to the codimension of  $V_G(T)$  in  $H$ ), that is,  $\text{codim}(T) := \dim(\otimes^n W)^G - \dim CV_G(T)$ . The dimension of  $(\otimes^n W)^G$  is  $m_1(n)$ , which has been computed in [16, Prop. 20] for the models listed in Example 3.4.

### 3.2 Smoothness at points of no evolution

Let  $T_n$  be the *claw  $n$ -tree*, that is, the tree with one inner vertex and  $n$  leaves. In what follows we prove that the variety corresponding to  $T_n$  is smooth at generic points of no evolution. In particular, it can be locally defined by a complete intersection.

Given a permutation subgroup  $G$  of  $\mathfrak{S}_\kappa$ , we denote by  $GL(\kappa)^G$  the group of  $G$ -equivariant  $\kappa \times \kappa$  invertible matrices. Clearly,  $GL(\kappa)^G$  defines an action on  $\text{Hom}_G(W, W)$  by  $(A, M) \rightarrow AM$ .

**Theorem 3.8** *The variety  $CV_G(T_n)$  is the Zariski closure of the orbit of  $\Psi_T^G(\mathbf{1}, \mathbf{I})$  under the group action of  $(GL(\kappa)^G)^n$ .*

PROOF. The conditions

1.  $\det A^e \neq 0$  for all  $e \in E(T)$ , and
2. all coordinates  $\pi$  are different from 0

define open sets in  $\text{Par}_G(T)$ . As  $\mathbf{I} \in GL(\kappa)^G$  and  $\mathbf{1} \in W^G$ , the intersection of these open sets is non-empty and a generic point  $(\pi, A^e) \in \text{Par}_G(T)$  satisfies both conditions. Let us fix one edge  $e_0$ . Notice that  $\text{diag}(\pi)A^{e_0}$  is invertible (as the coordinates of  $\pi$  are nonzero) and is  $G$ -invariant (as  $\pi \in W^G$ ). This means  $\text{diag}(\pi)A^{e_0} \in GL(\kappa)^G$ . Notice that  $\Psi_T^G(\pi, A^e) = \Psi_T^G(\mathbf{1}, \tilde{A}^e)$ , where  $\tilde{A}^e = A^e$  for  $e \neq e_0$  and  $\tilde{A}^{e_0} = \text{diag}(\pi)A^{e_0}$ . However,  $\Psi_T^G(\mathbf{1}, \tilde{A}^e) = (\tilde{A}^e) \cdot \Psi_T^G(\mathbf{1}, \mathbf{I})$ .  $\square$

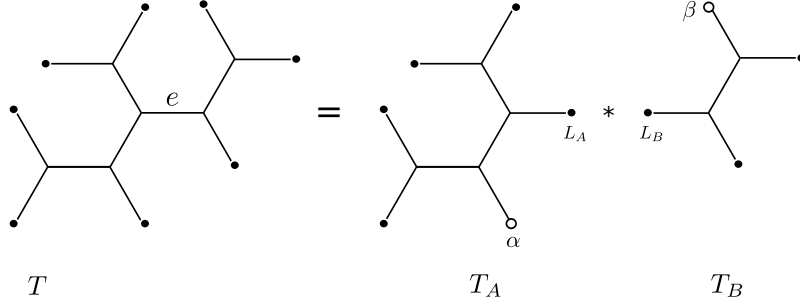


Figure 1: Decomposition of  $T$  into two subtrees  $T_A, T_B$ :  $T = T_A * T_B$ .

**Corollary 3.9** *If  $\sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} \mathbf{x} \otimes \dots \otimes \mathbf{x}$  satisfies  $\pi_{\mathbf{x}} \neq 0$  for all  $\mathbf{x}$ , then it is nonsingular. In particular, a generic point of no evolution of  $CV_G(T_n)$  and  $V_G(T_n)$  is nonsingular.*

PROOF. For  $CV_G(T_n)$  the statement follows directly from Theorem 3.8. Let us also notice that  $CV_G(T_n)$  is a cone, hence a point of  $V_G(T_n)$  is smooth if and only if it is smooth as a point of  $CV_G(T_n)$ .  $\square$

**Remark 3.10** When  $G$  acts on the basis of  $V$  transitively and freely then  $GL(V)^G$  is a torus. This is the case of so-called group-based models and, as follows from Theorem 3.8 the variety  $V_G(T_n)$  is toric [44].

## 4 Equations for the complete intersection

A *bipartition*  $A|B$  of the set of leaves is just a decomposition  $L(T) = A \cup B$ , where  $A$  and  $B$  are disjoint sets. Throughout the paper, we write  $a = |A|$ ,  $b = |B|$  and, to avoid trivialities, we will assume that  $a, b \geq 2$ . A bipartition is an *edge split* of  $T$  if it arises by removing one of the edges of  $T$ . If  $A \subset L(T)$ , we denote  $\otimes_{i \in A} W_i$  by  $W_A$ .

Given a tree  $T$ , in this section we proceed to construct equations that will define a complete intersection for  $CV_G(T)$ . We choose an internal edge  $e$  of  $T$ , which induces an edge split of the set of leaves  $L(T) = A \cup B$ . This allows us to view the tree  $T$  as the gluing of two trees  $T = T_A * T_B$  where  $L(T_A) = A \cup L_A$ ,  $L(T_B) = B \cup L_B$ , and  $L_A, L_B$  are the two vertices of the edge  $e$ , see Figure 1 and [6]. We assume that the leaves of  $T$  are ordered so that those in  $A$  appear in the first place, and those in  $B$  appear afterwards. We call  $\alpha \in L(T_A)$  the last leaf of  $A$  and  $\beta \in L(T_B)$  the first leaf of  $B$ .

A complete intersection for the variety  $CV_G(T)$  will be obtained by joining equations of a complete intersection for  $CV_G(T_A)$ , of a complete intersection for  $CV_G(T_B)$  and specific edge invariants. All results of this section (and the following) still hold if we replace the vector  $\mathbf{1}$  by any other  $G$ -invariant vector of  $W$ .

### 4.1 A basis linked to an edge split

In order to provide specific equations for the varieties associated to phylogenetic trees, we proceed to construct a basis  $\mathcal{B}_{A|B}$  of  $(\otimes^n W)^G$  related to a given edge split  $A|B$  as above. This basis shall be used to specify coordinates and provide the equations as polynomials in these specific

indeterminates. The important point is that this basis must be consistent with the decomposition

$$(\otimes^n W)^G \cong \oplus_{k=1}^t \text{Hom}_{\mathbb{C}}(\mathcal{F}_k(W_A^*), \mathcal{F}_k(W_B)). \quad (3)$$

given by Proposition 2.5.

We construct the desired basis of  $(\otimes^n W)^G$  compatible with (3) as follows.

**Algorithm to construct a basis of  $(\otimes^n W)^G$  linked to an edge split.**

1. Choose bases  $\{u_i^k\}_{i=1 \div m_k}$  of each  $\mathcal{F}_k(W)$ ,  $k = 1, \dots, t$ .
2. For each  $k = 1, \dots, t$ , the vectors  $u_{B,i}^k := u_i^k \otimes \mathbf{1}^{b-1}$  of  $\mathcal{F}_k(W_B)$ ,  $i = 1 \div m_k$ , are linearly independent. Indeed, the monomorphism  $W_B \xrightarrow{\cdot \otimes \mathbf{1}^{b-1}} W_B$  obtained by tensoring with a power of  $\mathbf{1}$  induces a monomorphism  $\mathcal{F}_k(W) \xrightarrow{\iota} \mathcal{F}_k(W_B)$ . We extend them to a basis  $\{u_{B,i}^k\}_{i=1 \div m_k(b)}$  of  $\mathcal{F}_k(W_B)$ .
3. We repeat step 2 for  $A$  to obtain a basis  $\{u_{A,i}^{k*}\}_{i=1 \div m_k(a)}$  of  $\mathcal{F}_{k*}(W_A)$  for each  $k$  (but now tensoring at the left,  $\mathbf{1}^{a-1} \otimes u_i^{k*}$ ).
4. Write  $S$  for the inverse of the isomorphism of Proposition 2.5. Its restrictions induce natural isomorphisms from

$$\mathcal{F}_{k*}(W_A) \otimes \mathcal{F}_k(W_B) \cong \text{Hom}_{\mathbb{C}}(\mathcal{F}_{k*}(W_A)^*, \mathcal{F}_k(W_B))$$

to

$$\text{Hom}_G(W_A^*[\chi_k], W_B[\chi_k]) \cong (W_A[\chi_{k*}] \otimes W_B[\chi_k])^G.$$

We call  $\mathcal{B}_{A|B}$  the desired basis  $\{S(u_{A,i}^{k*} \otimes u_{B,j}^k)\}_{i,j,k}$  of  $(\otimes^n W)^G$ .

From now on, we will denote by  $q_{ij}^k$  the coordinate corresponding to the basis vector  $S(u_{A,i}^{k*} \otimes u_{B,j}^k)$ .

**Remark 4.1** We describe here the isomorphism  $S$  mentioned above. Let  $f$  be the morphism in  $\text{Hom}_{\mathbb{C}}((\mathcal{F}_{k*}(W_A)^*, \mathcal{F}_k(W_B)))$  corresponding to  $u_{A,i}^{k*} \otimes u_{B,j}^k$  (this is,  $f(\omega) = \omega(u_{A,i}^{k*})u_{B,j}^k$ ). To present  $f$  as an element  $S(f) \in \text{Hom}_G(W_A^*[\chi_k], W_B[\chi_k])$  we proceed as follows:

1. Denote by  $\{(u_{A,i}^{k*})^*\}_i \subset \mathcal{F}_k(W_A^*)$  the dual basis for  $\{u_{A,i}^{k*}\}_i$ . Choose a subset  $H \subset G$  such that, for any  $i = 1 \div m_k(a)$ ,  $\{h(u_{A,i}^{k*})^*\}_{h \in H}$  is a basis of a subrepresentation in  $W_A^*[\chi_k]$  (necessarily isomorphic to  $N_k$ ).
2. Then,  $\{h(u_{A,i}^{k*})^*\}_{h \in H, i=1 \div m_k(a)}$  is a basis of  $W_A^*[\chi_k]$ .
3. We define  $S(f)(h(u_{A,i}^{k*})^*) = hu_{B,j}^k$  and  $S(f)(h(u_{A,l}^{k*})^*) = 0$  for  $l \neq i$ . This is the natural  $G$ -equivariant morphism associated to  $f$ .

The following statement claims that under stronger assumptions, the image of  $S(u_{A,i}^{k*} \otimes u_{B,j}^k)$  under the canonical isomorphism  $\text{Hom}_G(W_A^*[\chi_k], W_B[\chi_k]) \cong (W_A[\chi_{k*}] \otimes W_B[\chi_k])^G$  has the particularly nice form of the averaging operator. Namely,

**Lemma 4.2** *If we can choose  $H$  to be a subgroup of  $G$ , then*

$$S(u_{A,i}^{k*} \otimes u_{B,j}^k) = \frac{n_k}{|G|} \sum_{g \in G} (gu_{A,i}^{k*}) \otimes (gu_{B,j}^k),$$

where  $n_k$  is the cardinality of  $H$  (equal to the dimension of  $N_k$ ).

PROOF. First we notice that the kernel of the averaging operator always contains the kernel of  $S(u_{A,i}^{k*} \otimes u_{B,j}^k)$ . Moreover, the result of the averaging operator is a  $G$ -equivariant homomorphism. It remains to evaluate it on  $(u_{A,i}^{k*})^*$ . Notice that for each  $h \in H$ , we have  $G = \{g^{-1}h : g \in G\}$ . Hence we have the following equality

$$\sum_{g \in G} (u_{A,i}^k)^* (g u_{A,i}^{k*}) g u_{B,j}^k = \sum_{g \in G} (u_{A,i}^k)^* (g^{-1} h u_{A,i}^{k*}) g^{-1} h u_{B,j}^k.$$

Therefore, the right hand part of the equality in the lemma when evaluated at  $(u_{A,i}^{k*})^*$  is:

$$\begin{aligned} \frac{n_k}{|G|} \sum_{g \in G} (u_{A,i}^{k*})^* (g u_{A,i}^{k*}) g u_{B,j}^k &= \frac{1}{|G|} \sum_{g \in G, h \in H} g^{-1} \left( g (u_{A,i}^{k*})^* (h u_{A,i}^{k*}) h u_{B,j}^k \right) = \\ &= \frac{1}{|G|} \sum_{g \in G} g^{-1} \left( g (u_{A,i}^{k*})^* (h u_{A,i}^{k*}) h u_{B,j}^k \right). \end{aligned} \quad (4)$$

On the other hand we have

$$S(u_{A,i}^{k*} \otimes u_{B,j}^k) \left( (u_{A,i}^{k*})^* \right) = \frac{1}{|G|} \sum_{g \in G} g^{-1} \left( S(u_{A,i}^{k*} \otimes u_{B,j}^k) (g (u_{A,i}^{k*})^*) \right), \quad (5)$$

as  $S(u_{A,i}^{k*} \otimes u_{B,j}^k)$  is equivariant. So far we did not use the fact that  $H$  is a subgroup. However, in such case  $H u_{A,i}^{k*}$  is the dual basis to  $H((u_{A,i}^{k*})^*)$ , i.e.  $h(u_{A,i}^{k*})^* = (h u_{A,i}^{k*})^*$ . In particular,

$$g(u_{A,i}^{k*})^* = \sum_{h \in H} g(u_{A,i}^{k*})^* (h u_{A,i}^{k*}) (h u_{A,i}^{k*})^* = \sum_{h \in H} g(u_{A,i}^{k*})^* (h u_{A,i}^{k*}) h(u_{A,i}^{k*})^*.$$

Substituting this in (5) clearly yields the expression in (4), by point 3 in Remark 4.1.  $\square$

#### 4.1.1 Some examples

For the models of Example 3.4, we consider the Fourier basis of the space  $W$ , defined as  $\underline{\Sigma} := \{\underline{\mathbf{A}}, \underline{\mathbf{C}}, \underline{\mathbf{G}}, \underline{\mathbf{T}}\}$  of  $W$  where

$$\underline{\mathbf{A}} = \mathbf{A} + \mathbf{C} + \mathbf{G} + \mathbf{T}; \quad \underline{\mathbf{C}} = \mathbf{A} + \mathbf{C} - \mathbf{G} - \mathbf{T}; \quad \underline{\mathbf{G}} = \mathbf{A} - \mathbf{C} + \mathbf{G} - \mathbf{T}; \quad \underline{\mathbf{T}} = \mathbf{A} - \mathbf{C} - \mathbf{G} + \mathbf{T}.$$

Notice that  $\underline{\mathbf{A}}$  equals the vector  $\mathbf{1}$  introduced above and is invariant under the action of any permutation of  $\mathfrak{S}_4$ . Notice also that the permutation groups associated to these models have only real characters; so for every irreducible representation, it holds  $k^* = k$ . Throughout this section, we adopt the following notation: given  $\underline{\mathbf{X}}_i \in \underline{\Sigma}$ , we write  $\underline{\mathbf{X}}_1 \dots \underline{\mathbf{X}}_m$  for the tensor  $\underline{\mathbf{X}}_1 \otimes \dots \otimes \underline{\mathbf{X}}_m \in \otimes^m W$ .

**Example 4.3** To illustrate the first step of the algorithm of the previous section, we proceed to obtain basis of each space  $\mathcal{F}_k(W)$  when the group  $G$  is chosen according to some of the models of Example 3.4. All these models satisfy the following property:

(\*) the isotypic components of  $W$  can be spanned by some elements of the Fourier basis above.

Namely,

1.  $G = \{1\}$  (GMM). In this case, the only representation is the *identity* representation. We can take  $u_1^1 = \underline{\mathbf{A}}$ ,  $u_2^1 = \underline{\mathbf{C}}$ ,  $u_3^1 = \underline{\mathbf{G}}$ ,  $u_4^1 = \underline{\mathbf{T}}$ , which form a basis of  $\mathcal{F}_1(W) = W$ .

$\Omega_{\mathfrak{S}_4}$	id	(AC)	(ACG)	(ACGT)	(AC)(GT)
$\chi_1$	1	1	1	1	1
$\chi_2$	1	-1	1	-1	1
$\chi_3$	2	0	-1	0	2
$\chi_4$	3	1	0	-1	-1
$\chi_5$	3	-1	0	1	-1
$\chi$	4	2	1	0	0

$\Omega_{\mathbb{Z}_2}$	id	(AT)(CG)
$\chi_1$	1	1
$\chi_2$	1	-1
$\chi$	4	0

Table 1: Character tables of the groups  $\mathfrak{S}_4$  and  $G = \langle(\text{AT})(\text{CG})\rangle$ . The character  $\chi$  corresponds to the permutation representation of the group on the space  $W$ .

2.  $G = \langle(\text{AT})(\text{CG})\rangle \cong \mathbb{Z}_2$  (strand symmetric model). There are two irreducible representations: the *identity* and the *sign* representations. By taking  $u_1^1 = \underline{\mathbf{A}}$ ,  $u_2^1 = \underline{\mathbf{T}}$ ,  $u_1^2 = \underline{\mathbf{C}}$ ,  $u_2^2 = \underline{\mathbf{G}}$ , we have that  $\{u_1^1, u_2^1\}$  and  $\{u_1^2, u_2^2\}$  are basis of  $\mathcal{F}_1(W) = W[\chi_1]$  and  $\mathcal{F}_2(W) = W[\chi_2]$ , respectively.
3.  $G = \langle(\text{AC})(\text{GT}), (\text{AG})(\text{CT})\rangle \cong \mathbb{Z}_2 \times \mathbb{Z}_2$  (Kimura 3-parameter). There are four irreducible representations, each with dimension one (since  $G$  is abelian). Then, we can take  $u_1^1 = \underline{\mathbf{A}}$ ,  $u_1^2 = \underline{\mathbf{C}}$ ,  $u_1^3 = \underline{\mathbf{G}}$  and  $u_1^4 = \underline{\mathbf{T}}$ , so that each  $\mathcal{F}_k(W)$  is spanned by the corresponding  $u_1^k$ .
4.  $G = \langle(\text{ACGT}), (\text{AG})\rangle$  (Kimura 2-parameter). There are two irreducible representations for  $G$  with dimension 1. Taking  $u_1^1 = \underline{\mathbf{A}}$ ,  $u_1^2 = \underline{\mathbf{G}}$ , we obtain  $\mathcal{F}_k(W) = \langle u_1^k \rangle$ , for  $k = 1, 2$ . There is still a 2-dimensional irreducible representation; we can take  $u_1^3 = \underline{\mathbf{C}}$  to get a basis of the corresponding space  $\mathcal{F}_3(W)$  (a different possibility would be to take  $u_1^3 = \underline{\mathbf{T}}$ ).
5.  $G = \mathfrak{S}_4$  (Jukes-Cantor). There are five irreducible representation, but only two of them appear in the Maschke decomposition of  $W$ : the identity representation and one 3-dimensional representation with character  $\chi_4$  in Table 1. By taking  $u_1^1 = \underline{\mathbf{A}}$  and  $u_1^4 = \underline{\mathbf{C}}$ , we obtain bases for the spaces  $\mathcal{F}_1(W)$  and  $\mathcal{F}_4(W)$ .

**Remark 4.4** There exist equivariant models that do not satisfy the property (\*) above. For example, if  $G = \langle(\text{AC})\rangle \cong \mathbb{Z}_2$ , there are two irreducible representations  $\chi_1, \chi_2$  and the Maschke decomposition of  $W$  becomes  $W = W[\chi_1] \oplus W[\chi_2]$ , where  $W[\chi_1] = \langle \underline{\mathbf{A}} \rangle \oplus \langle \underline{\mathbf{C}} \rangle \oplus \langle \underline{\mathbf{G}} + \underline{\mathbf{T}} \rangle$  and  $W[\chi_2] = \langle \underline{\mathbf{G}} - \underline{\mathbf{T}} \rangle$ .

**Example 4.5 A basis linked to a bipartition for the strand symmetric model.** Take  $G = \langle(\text{AT})(\text{CG})\rangle \cong \mathbb{Z}_2$ , so we deal with the strand symmetric model. The character table of  $G$  is shown in Table 1.

The permutation representation of  $G$  decomposes as  $\chi = 2\chi_1 + 2\chi_2$ , and  $W = W[\chi_1] \oplus W[\chi_2]$ , with  $W[\chi_1] = \langle \underline{\mathbf{A}}, \underline{\mathbf{T}} \rangle$  and  $W[\chi_2] = \langle \underline{\mathbf{C}}, \underline{\mathbf{G}} \rangle$ .

On the tree 12|34, we consider the edge split  $A = \{1, 2\}$ ,  $B = \{3, 4\}$ ,  $\alpha = 2$ ,  $\beta = 3$ . The vectors  $u_1^1 = \underline{\mathbf{A}}$ ,  $u_2^1 = \underline{\mathbf{T}}$ ,  $u_1^2 = \underline{\mathbf{C}}$ ,  $u_2^2 = \underline{\mathbf{G}}$  regarded as vectors in  $W_\alpha$  induce tensors in  $\mathcal{F}_1(W_A)$  and  $\mathcal{F}_2(W_A)$ , just by tensoring with  $\mathbf{1} = \underline{\mathbf{A}}$  on the left:  $u_{A,1}^1 = \underline{\mathbf{AA}}$ ,  $u_{A,2}^1 = \underline{\mathbf{AT}}$ ,  $u_{A,1}^2 = \underline{\mathbf{AC}}$  and  $u_{A,2}^2 = \underline{\mathbf{AG}}$ . We extend these tensors to a basis of  $\mathcal{F}_1(W_A)$  and  $\mathcal{F}_2(W_A)$  with

$$\begin{aligned}
u_{A,3}^1 &= \underline{\mathbf{TA}}, \quad u_{A,4}^1 = \underline{\mathbf{TT}}, \quad u_{A,5}^1 = \underline{\mathbf{CC}}, \quad u_{A,6}^1 = \underline{\mathbf{CG}}, \quad u_{A,7}^1 = \underline{\mathbf{GC}}, \quad u_{A,8}^1 = \underline{\mathbf{GG}}, \\
u_{A,3}^2 &= \underline{\mathbf{TC}}, \quad u_{A,4}^2 = \underline{\mathbf{TG}}, \quad u_{A,5}^2 = \underline{\mathbf{CA}}, \quad u_{A,6}^2 = \underline{\mathbf{CT}}, \quad u_{A,7}^2 = \underline{\mathbf{GA}}, \quad u_{A,8}^2 = \underline{\mathbf{GT}}.
\end{aligned}$$

We proceed similarly for  $B$ , and then we construct the basis  $\{S(u_{A,i}^k \otimes u_{B,j}^k)\}_{k,i,j}$  of  $(\otimes^4 W)^G$ . As the two irreducible representations of  $G$  are 1-dimensional, the  $S$  operator has no effect and  $\{u_{A,i}^k \otimes u_{B,j}^k\}_{k,i,j}$  is already a basis linked to  $A|B$ .

## 4.2 Explicit Edge invariants

Once an edge split  $A|B$  of the tree topology  $T$  is given, edge invariants associated to it arise as restrictions on the rank of some matrices  $M_k$ ,  $k = 1, \dots, t$ . Our goal here is to explain how these matrices arise, and investigate how these rank restrictions look like.

The decomposition (3) allows us to understand any tensor  $p \in (\otimes^n W)^G$  as a collection  $(g_p^1, g_p^2, \dots, g_p^t)$ , where each  $g_p^k : \mathcal{F}_k(W_A^*) \rightarrow \mathcal{F}_k(W_B)$  is a linear map.

**Definition 4.6** [Thin flattening] The collection of linear maps constructed above is referred to as the *thin flattening of  $p$  relative to the bipartition  $A|B$* :  $\text{Tflat}_{A|B}(p) = (g_p^1, g_p^2, \dots, g_p^t)$ .

The main result of [14] claims that if  $p$  is a (general) point in  $CV_G(T)$ , then the bipartition  $A|B$  is an edge split in  $T$  if and only if

$$\text{rank } g_p^k \leq m_k, \quad \text{for every } k = 1, \dots, t.$$

The  $(m_k + 1) \times (m_k + 1)$  minors of matrices representing  $g_k$  are usually known as *edge invariants*.

We consider the basis  $\mathcal{B}_{A|B}$  of  $(\otimes^n W)^G$  linked to the edge split  $A|B$  constructed in section 4.1. As we have fixed bases  $\{u_{A,i}^{k*}\}_i$  of  $\mathcal{F}_{k*}(W_A)$  and  $\{u_{B,j}^k\}_j$  of  $\mathcal{F}_k(W_B)$ , each tensor in  $(\otimes^n W)^G$  naturally induces matrices  $M_k$  representing the morphisms  $g_p^k \in \text{Hom}_{\mathbb{C}}(\mathcal{F}_{k*}(W_A), \mathcal{F}_k(W_B))$  of the thin flattening. Each rank restriction for  $M_k$ , is an equation on the coordinates  $q_{i,j}^k$  introduced in section 4.1.

In order to obtain a complete intersection, we shall now choose specific minors of order  $m_k + 1$  in the matrices  $M_k$ . The basis we constructed for  $\mathcal{F}_k(W_B)$  (respectively  $\mathcal{F}_{k*}(W_A)$ ) has a distinguished set of  $m_k$  (respectively  $m_{k*}$ ) elements, namely the first  $m_k$  (resp.  $m_{k*}$ ) elements. We call  $M_k^0$  the submatrix of  $M_k$  corresponding to these elements.

We choose only the  $(m_{k*} + 1) \times (m_k + 1)$  minors of  $M_k$  that contain the distinguished  $m_{k*} \times m_k$ -submatrix  $M_k^0$ . As in our setting we have  $k = k^*$ , we observe that  $M_k^0$  is a square matrix.

For the purpose of the next section, we need to write these minors in terms of the determinant of  $M_k^0$ ,  $\Delta_k(p) = \det M_k^0$ . Note that  $M_k^0$  is the upper left  $m_k \times m_k$  submatrix of  $M_k$  so that  $\det M_k^0$  is a polynomial in indeterminates  $q_{ij}^k$  for  $1 \leq i \leq m_{k*}$ ,  $1 \leq j \leq m_k$ .

We note by  $E_{i,j}^k$  the minor containing  $M_k^0$ , the row indexed by  $u_{A,i}^{k*}$  and the column indexed by  $u_{B,j}^k$ . Then the minors  $E_{i,j}^k$  containing  $M_k^0$ ,  $k = 1 \div t$ ,  $i = m_{k*} + 1 \div m_{k*}(a)$ ,  $j = m_k + 1 \div m_k(b)$ , can be written as

$$E_{ij}^k = q_{ij}^k \Delta_k(p) + \sum_{s=1}^{m_k} (-1)^{j+s} q_{sj}^k \Delta_k^{(s)}(p) \quad (6)$$

where  $\Delta_k^{(s)}(p)$  is the determinant of the matrix obtained by removing the  $s$ -th row of  $M_k^0$  and adding the first  $m_k$  entries of the  $i$ -th row of  $M_k$ . The set of equations  $E_{i,j}^k = 0$  (which are a particular subset of the edge invariants for  $A|B$ ) will be denoted by  $\mathbf{eq}_{A|B}$ . There are

$$N_{A|B} := \sum_k (m_{k*}(a) - m_{k*})(m_k(b) - m_k)$$

such minors.

**Remark 4.7** Notice that the cardinality  $N_{A|B}$  of this set depends only on  $a = |A|$  and  $b = |B|$ , and not of the particular choice of leaves in  $A$  or  $B$ . Moreover, by Proposition 2.5, we have  $N_{A|B} = m_1(n) - m_1(a+1) - m_1(b+1) + m_1(2)$ . For the models of Example 3.4, explicit formulas for  $m_1(s)$  are given in [16, Prop. 5]. From these formulas, it is easy to see that  $N_{A|B}$  grows exponentially with  $n$ . Therefore, the list of local phylogenetic invariants we give in the following section has exponential cardinality in  $n$ , which would make it useless for practical applications when  $n$  is big. However, it is well known that rank conditions do not need to be checked directly by evaluating minors; they can be checked by using the singular values of the matrix instead (this is the approach followed in [28, 15]). Thus, for the reader interested in applying the local phylogenetic invariants provided in this paper, we suggest using singular values instead of  $\mathbf{eq}_{A|B}$ .

The following result, which is needed in the next section, easily follows from (6).

**Lemma 4.8** *For any  $k$  and  $m_{k^*} < i', j', i, j \leq m_{k^*}(a)$ , we have*

$$\frac{\partial E_{ij}^k}{\partial q_{i'j'}^k} = \begin{cases} \Delta_k(p), & \text{if } (i, j) = (i', j'); \\ 0, & \text{otherwise.} \end{cases}$$

### 4.3 Equations from $CV_G(T_A)$ and $CV_G(T_B)$

The following result is essentially well-known (see Lemma 1 of [30], or [5]). However, we prove it in our setting.

**Lemma 4.9** *Let  $T$  be a tree and let  $L$  be one of its leaves. Let  $T'$  be the subtree of  $T$  that has the same vertices, apart from  $L$  (Fig. 2). The following contraction map*

$$\begin{aligned} f_L : \otimes^n W &= \bigotimes_{l \in L(T)} W_l \rightarrow \bigotimes_{l \in L(T')} W_l \\ \otimes_{l \in L(T)} v_l &\mapsto \mathbf{1} \cdot v_L \left( \otimes_{l \in L(T')} v_l \right) \end{aligned}$$

*satisfies  $f_L(\text{Im } \Psi_T^G) = \text{Im } \Psi_{T'}^G$  and, as a consequence,  $\overline{f_L(CV_G(T))} = CV_G(T')$ .*

In stochastic terms, this map is called the *marginalization over the random variable at  $L$* .

**PROOF.** The map  $f_L$  is induced by the multilinear map

$$\begin{aligned} \prod_{l \in L(T)} W_l &\rightarrow \bigotimes_{l \in L(T')} W_l \\ (v_l)_{l \in L(T)} &\mapsto \mathbf{1} \cdot v_L \left( \otimes_{l \in L(T')} v_l \right) \end{aligned}$$

and therefore  $f_L$  is well defined (by the universal property of tensor products).

Without loss of generality, we can assume that the interior node  $m$  adjacent to  $L$  in  $T$  has degree  $\geq 3$  (indeed, if it had degree two, then  $CV_G(T)$  would be isomorphic to the variety associated to the tree with this vertex removed and two adjacent edges joined into a single edge).

We call  $v_1, \dots, v_t$  ( $t \geq 3$ ) the vertices adjacent to  $m$  and we set  $v_1 = L$ . We root the tree  $T$  at  $v_t$  and call  $e(m, i)$  the edges from  $m$  to  $v_i$ ,  $i = 1 \dots, t-1$  (see figure 2).

Let  $\mathcal{P} = \left( \pi, (A^e)_{e \in E(T)} \right)$  be a point in  $\text{Par}_G(T)$  (rooted at  $v_t$ ). For the edge  $e(m, 2)$  from  $m$  to  $v_2$ , we consider a new matrix  $B^{e(m, 2)} := D A^{e(m, 2)}$  where  $D$  is the diagonal matrix  $\text{diag}(A^{e(m, L)} \mathbf{1})$  formed by the entries of  $A^{e(m, L)} \mathbf{1}$ . Since  $A^{e(m, 2)}$  is  $G$ -equivariant, the vector  $A^{e(m, 2)} \mathbf{1}$  is  $G$ -invariant, and  $D$  is  $G$ -equivariant again. It follows from this that the new matrix  $B^{e(m, 2)}$  is  $G$ -equivariant. For all other edges of  $T'$ , take  $B^e = A^e$ . It is not difficult to check that

$$f_L(\Psi_T^G(\mathcal{P})) = \Psi_{T'}^G \left( \pi, (B^e)_{e \in E(T')} \right).$$

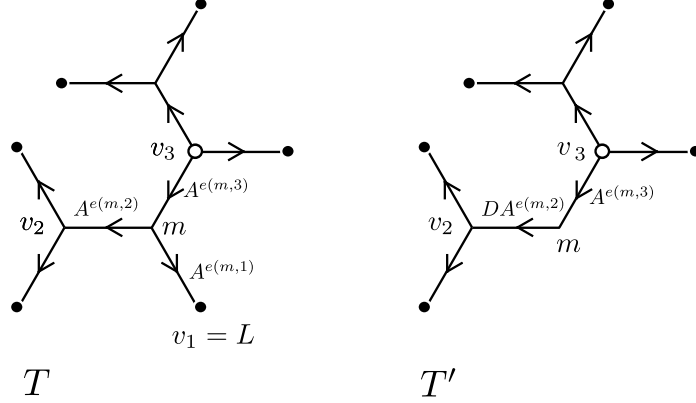


Figure 2: . Illustration of the proof of Lemma 4.9. The tree  $T'$  is obtained by taking the leaf  $v_1 = L$  off the tree  $T$ .

Therefore  $f_L(\text{Im } \Psi_T^G) \subset \text{Im } \Psi_{T'}^G$ . The other inclusion  $f_L(\text{Im } \Psi_{T'}^G) \supseteq \text{Im } \Psi_T^G$  follows easily by adding the identity matrix at the edge  $e(m, 1)$ .

The equality of the parametrized part of the varieties implies equality in the closures, hence  $\overline{f_L(CV_G(T))} = CV_G(T')$ .  $\square$

By successive applications of Lemma 4.9 to all leaves in  $A \setminus \{\alpha\} \subset L(T)$  (that is, marginalizing over all leaves in  $A \setminus \{\alpha\}$ ) we obtain a map

$$f_{A \setminus \alpha} : \bigotimes_{l \in L(T)} W_l \longrightarrow \bigotimes_{l \in L(T_B)} W_l$$

that sends the variety  $CV_G(T)$  to  $CV_G(T_B)$ .

In order to induce equations from  $T_A$ ,  $T_B$  to  $T$ , below we translate this map in terms of the corresponding affine coordinate rings. We do not explicitly write indeterminates nor coordinates because the above map  $f_{A \setminus \alpha}$  is basis independent. This fact will play an important role in the proof of the main result in the next section.

The map  $f_{A \setminus \{\alpha\}}$  above is dual to the map

$$\bigotimes_{l \in L(T_B)} W_l^* \rightarrow \bigotimes_{l \in L(T)} W_l^*$$

that maps  $t$  to  $\mathbf{1}^{a-1} \otimes t$  if the leaf  $L_B$  of  $T_B$  is identified with the leaf  $\alpha$  of  $T$ , so  $f_{A \setminus \{\alpha\}}^*$  is the map corresponding to  $f_{A \setminus \{\alpha\}}$  in terms of coordinates. Moreover, both maps restrict to  $G$ -invariant vectors. Summing up we have:

**Corollary 4.10** *Any equation vanishing on  $CV_G(T_B)$  extends to an equation vanishing on  $CV_G(T)$  via the map:*

$$f_{A \setminus \alpha}^* : \left( \bigotimes_{l \in L(T_B)} W_l^* \right)^G \xrightarrow{t} \left( \bigotimes_{l \in L(T)} W_l^* \right)^G \xrightarrow{\mapsto} \mathbf{1}^{a-1} \otimes t$$

where the leaf  $L_B$  of  $T_B$  is identified with the leaf  $\alpha$  of  $T$ .



Similarly, for any subsets  $R \subsetneq S \subset L(T)$  of leaves of  $T$  we have  $f_S^*$  is the map

$$\begin{aligned} f_S^* : \bigotimes_{l \in R \setminus S} W_l^* &\longrightarrow \bigotimes_{l \in R} W_l^* \\ \bigotimes_{l \in R \setminus S} v_l &\mapsto \bigotimes_{l \in R} w_l, \end{aligned}$$

where  $w_l = v_l$  if  $l \in R \setminus S$  and  $w_l = \mathbf{1}$  if  $l \in S$ .

**Remark 4.11** It is convenient to write the equations of  $T_B$  in the basis related to a bipartition  $L_B|B$ . In this way, the extension of a coordinate as defined in Corollary 4.10 gives rise to a coordinate that is already in the basis  $\mathcal{B}_{A|B}$  of  $(\otimes W^n)^G$  (indeed, as  $\mathbf{1}$  is  $G$ -invariant, the operator (4.1) does not affect it and it is easy to check that the extended basis are elements of  $\mathcal{B}_{A|B}$ ).

## 5 The main result

Given a phylogenetic tree under an equivariant model  $\mathcal{M}_G$ , the goal of this section is to construct a complete intersection for  $CV_G(T)$  on a neighborhood of a point of no evolution. This will be done by using induction on the number of leaves of the tree.

Let  $T$  be a tree with at least one interior edge and leaves  $L(T) = \{l_1, \dots, l_n\}$ . Reordering the set of leaves (if needed) we can assume that there exists a node with children  $\{l_{n-l}, \dots, l_n\}$ ,  $l \geq 1$ . We take the edge split  $A|B$  given by  $A = \{l_1, \dots, l_{n-l-1}\}$ ,  $B = \{l_{n-l}, \dots, l_n\}$ . Write  $e$  for the interior edge of  $T$  associated with this split. Keeping the notation already used throughout the paper,  $T_A$  has leaves  $A \cup \{L_A\}$ ,  $T_B$  has leaves  $B \cup \{L_B\}$  (where  $L_A, L_B$  are defined as in figure 1). The variety associated to  $T_A$  is the closure of the image of the polynomial map

$$\Psi_{T_A} : \text{Par}_G(T_A) \rightarrow \otimes^{n-l} W^G.$$

and the variety  $CV_G(T_B) \subset \otimes^{l+2} W^G$  is the closure of the image of

$$\Psi_{T_B} : \text{Par}_G(T_B) \rightarrow \otimes^{l+2} W^G.$$

**Lemma 5.1** *Given a leaf  $L$  in  $T$ , the image of a point of no evolution  $\pi_n \in \otimes^n W$  under the map  $f_L$  is a point of no evolution in  $\otimes^{n-1} W$ .*

PROOF. If  $\pi_n = \sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} \mathbf{x} \otimes \dots \otimes \mathbf{x}$ , then its image under the map  $f_L$  is

$$f_L(\pi_n) = \sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} (\mathbf{1} \cdot \mathbf{x})^a (\mathbf{x} \otimes \overset{n-a}{\mathbf{x}}) = \sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} \mathbf{x} \otimes \overset{n-a}{\mathbf{x}}.$$

As  $\pi_n$  was invariant by the action of  $G$ , so is  $f_L(\pi_n)$  and therefore it is a point of no evolution in  $\otimes^{n-a} W$ .  $\square$

By successively applying the above lemma and lemma 4.9 we obtain points of no evolution in  $CV_G(T_A)$  and in  $CV_G(T_B)$  from a point of no evolution  $\pi_n$  in  $CV_G(T)$ . This shall allow us to apply an induction argument.

Let  $T_d$  be the claw tree with  $d$  leaves (claw  $d$ -tree) evolving under  $\mathcal{M}_G$  and  $L(T_d) = \{x_0, \dots, x_{d-1}\}$ , and let  $\mathbf{eq}_{T_d} := \{h_1, h_2, \dots, h_{\text{codim}(CV_G(T_d))}\}$  be a set of equations of a complete intersection that defines  $CV_G(T_d) \subset (\otimes^d W)^G$  on an open subset containing general points of no evolution. As we already proved, general points of no evolution are smooth by Theorem 3.8. In particular, the variety is locally a complete intersection, which guaranties the existence of  $\mathbf{eq}_{T_d}$ . Before proceeding with induction, we need the following assumption about the equivariant model  $\mathcal{M}_G$  on the claw tree with  $d$  leaves, which shall be checked for every particular equivariant model and every  $d$  equal

to a degree of one of the interior nodes of  $T$ . For the GMM, the strand symmetric model, and the Jukes-Cantor model, we prove in section 6 that this assumption holds for the tripod (and hence our result is valid for trivalent trees evolving on these models). A local complete intersection for the Kimura 3-parameter model for trivalent trees was already given in [13].

**Claw  $d$ -tree hypothesis 5.2** We write equations  $\mathbf{eq}_{T_d}$  on a basis of type  $\mathcal{B}_{x_0|\{x_1, \dots, x_{d-1}\}}$  following subsection 4.2. The jacobian of these new equations  $\mathbf{eq}_B$ , which we denote as  $\mathbf{J}_{x_0|x_1, \dots, x_{d-1}}(T_d)$ , has rank equal to  $\text{codim}(CV_G(T_d))$  at any general point of no evolution. We denote by  $\mathbf{J}_{x_0|x_1, \dots, x_{d-1}}^*(T_d)$  the matrix obtained from  $\mathbf{J}_{x_0|x_1, \dots, x_{d-1}}(T_d)$  by removing the columns corresponding to  $S(u_{x_0, i}^k \otimes u_{x_1, \dots, x_{d-1}, j}^k)$  for  $k = 1 \div t$  and  $i, j = 1 \div m_k$ . We say that the equivariant model  $\mathcal{M}_G$  satisfies the *claw  $d$ -tree hypothesis* if

$$\text{rank } \mathbf{J}_{x_0|x_1, \dots, x_{d-1}}^*(T_d) = \text{codim}(CV_G(T_d)), \quad (7)$$

whenever this matrix is evaluated at a generic point of no evolution.

**Induction hypothesis.** We will use the following induction hypothesis:

(\*) There is a set of equations  $\mathbf{eq}_{T_A} = \{g_1, g_2, \dots, g_{\text{codim}(CV_G(T_A))}\}$  that defines the variety  $CV_G(T_A) \subset (\otimes^{n-l} W)^G$  scheme theoretically on an open subset containing general points of no evolution.

By Corollary 4.10, the map  $\tau \mapsto \tau \otimes \mathbf{1}^l$  induces new equations for  $CV_G(T)$  from  $\mathbf{eq}_{T_A}$ . These equations shall be written in the coordinates  $q_{i,j}^k$  corresponding to the basis  $\mathcal{B}_{A|B}$  linked to the edge split  $A|B$  and shall be called  $\mathbf{eq}_A = \{f_1^A, \dots, f_{\text{codim}(CV_G(T_A))}^A\} \subset \mathbb{C}[q_{i,j}^k]$ .

As above, by Corollary 4.10, the map  $\tau \mapsto \mathbf{1}^{n-l-2} \otimes \tau$  induces new equations

$$\mathbf{eq}_B = \{f_1^B, \dots, f_{\text{codim}(CV_G(T_B))}^B\} \subset \mathbb{C}[q_{i,j}^k]$$

for  $CV_G(T)$  from the set of equations  $\mathbf{eq}_{T_B}$  of the underlying model assumption.

Besides, we still need to consider the set of polynomials coming from the edge split.

**Edge invariants.** As in subsection 4.2, for each  $k = 1, \dots, t$ , write  $M_k$  for the  $m_k(n-l-1) \times m_k(l+1)$ -matrix with rows indexed by the  $u_{A,i}^k$ ,  $i = 1, \dots, m_k(n-l-1)$ , columns indexed by  $u_{B,j}^k$ ,  $j = 1, \dots, m_k(l+1)$ , and whose  $(i, j)$ -entry is the coordinate  $q_{i,j}^k$ . For each of these matrices, take the set of all the  $(m_k+1) \times (m_k+1)$ -minors containing the sub matrix  $M_k^0$  defined in Section 4.2, with rows and columns indexed by  $\{u_{A,i}^k\}_{i=1, \dots, m_k}$  and  $\{u_{B,j}^k\}_{j=1, \dots, m_k}$ , respectively. We obtain  $N_{A|B}$  polynomials in  $\mathbb{C}[q_{i,j}^k]$  of the form (6).

**Lemma 5.3** *We have  $\text{codim}(CV_G(T_A)) + \text{codim}(CV_G(T_B)) + N_{A|B} = \text{codim}(CV_G(T))$ .*

PROOF.

We assume that  $T$  has no vertices of degree 2, as such nodes can be removed. By Theorem 3.5 and Remark 4.7, we have

$$\begin{aligned} & \text{codim}(CV_G(T_A)) + \text{codim}(CV_G(T_B)) + N_{A|B} = \\ & (m_1(l+2) - (|E(T)| - (n-l-1))(m_1(2) - m_1) - m_1) + \\ & (m_1(n) - m_1(n-l) - m_1(l+2) + m_1(2)) + \\ & + (m_1(n-l) - (n-l)(m_1(2) - m_1) - m_1). \end{aligned}$$

The sum above equals

$$m_1(n) - |E(T)|(m_1(2) - m_1) - m_1 = \text{codim}(CV_G(T)).$$

□

**Theorem 5.4** *Let  $T$  be a phylogenetic tree on  $n$  leaves,  $n \geq 3$ ; let  $D$  be the set of degrees of its interior nodes, and assume that  $d \geq 3$  for any  $d \in D$ . Let  $\mathcal{M}_G$  be an equivariant model that satisfies the claw  $d$ -tree hypothesis for any  $d \in D$ . The set of equations  $\mathbf{eq}_T := \mathbf{eq}_A \cup \mathbf{eq}_B \cup \mathbf{eq}_{A|B}$  defines the variety  $CV_G(T)$  scheme theoretically on an open subset that contains general points of no evolution.*

PROOF. We proceed by induction on the number of leaves. The first step is  $n = 3$ , which is covered by the claw  $d$ -tree assumption for  $d = 3$ . We assume thus  $n > 3$  and that  $T$  has at least one interior edge which splits the leaves  $L(T) = \{l_1, \dots, l_n\}$  into two sets  $A = \{l_1, \dots, l_{n-l-1}\}$  and  $B = \{l_{n-l}, \dots, l_n\}$  (reordering leaves if necessary). Consider the trees  $T_A$  and  $T_B$  as defined in the beginning of this section. Note that we are able to use the induction hypothesis stated above because the set of degrees for the interior nodes of the tree  $T_A$  is included in  $D$ . By Lemma 5.3, we know that

$$|\mathbf{eq}_A| + |\mathbf{eq}_B| + |\mathbf{eq}_{A|B}| = \text{codim}(CV_G(T)),$$

that is, the number of equations equals the codimension of the variety. We already know that  $\mathbf{eq}_T$  are equations satisfied by all points in  $CV_G(T)$ .

Let  $V'$  be the variety defined by  $\mathbf{eq}_T$ . Now, consider the jacobian matrix  $\mathbf{J}_{A|B}(T)$  obtained by taking the partial derivatives of the polynomials in  $\mathbf{eq}_{A|B}$  with respect to the coordinates  $q_{i,j}^k$  of  $(\otimes^n W)^G$ . We claim that the rank of this matrix at a generic point of no evolution  $\pi_n$  is maximal. From this, we will deduce that  $V'$  is non-singular in a neighborhood  $U$  of  $\pi_n$ . Since  $V'$  and  $CV_G(T)$  have the same dimension, it follows that both varieties are equal in  $U$ .

By reordering the columns of the jacobian matrix if necessary, we may assume that columns are indexed as follows:

- the first  $m_1(n-l) - m_1(2)$  columns are indexed by  $q_{i,j}^k$  with  $i = m_k + 1 \div m_k(n-l-1)$ ,  $j = 1 \div m_k$ ,  $k = 1 \div t$ ;
- then,  $m_1(2)$  columns indexed by  $q_{i,j}^k$  with  $i = 1 \div m_k$ ,  $j = 1 \div m_k$ ,  $k = 1 \div t$ ;
- then,  $m_1(l+2) - m_1(2)$  columns indexed by  $q_{i,j}^k$  with  $i = 1 \div m_k$ ,  $j = m_k + 1 \div m_k(l+1)$ ,  $k = 1 \div t$ ;
- the remaining columns correspond to  $i = m_k + 1 \div m_k(n-l-1)$ ,  $j = m_k + 1 \div m_k(l+1)$ ,  $k = 1 \div t$ .

Notice that the equations in  $\mathbf{eq}_A$  only have coordinates in the first  $m_1(n-l)$  columns and  $\mathbf{eq}_B$  only in the middle  $m_1(l+2)$  columns. With this ordering, the jacobian matrix has the form

$$\mathbf{J}_{A|B}(T) = \begin{pmatrix} \boxed{\mathbf{J}_{A|L_A}(T_A)} & 0 & 0 \\ * & \boxed{\mathbf{J}_{L_B|B}(T_B)} & 0 \\ \boxed{\frac{\partial}{\partial q_{i,j}^k} E_{i,j}^k} \end{pmatrix}$$

where:

1. The first block  $\mathbf{J}_{A|L_A}(T_A)$  has  $\text{codim}(CV_{T_A})$  rows, and  $m_1(n-l)$  columns indexed by the coordinates in  $(\otimes^{n-l}W)^G$  extended to  $(\otimes^n W)^G$ .
2. The second block  $\mathbf{J}_{L_B|B}(T_B)$  has  $\text{codim}(CV_{T_B})$  rows, and  $m_1(l+2)$  columns indexed by the coordinates in  $(\otimes^{l+2}W)^G$  extended to  $(\otimes^n W)^G$ .

Notice that the first two blocks share the columns indexed by the coordinates  $\{q_{i,j}^k\}_{1 \leq i,j \leq m_k}$ .

3. The third block  $(\frac{\partial}{\partial q_{i,j}^k} E_{i,j}^k)$  has  $N_{n-l-1|l+1}$  rows, indexed by the equations of  $\mathfrak{eq}_{A|B}^n$ , and  $m_1(n)$  columns, indexed by all the coordinates above:  $\{q_{i,j}^k\}_{1 \leq i \leq m_k(l+1), 1 \leq j \leq m_k(n-l-1), 1 \leq k \leq t}$ .

From now on, these coordinates refer to a generic point of no evolution  $\pi_n$ .

We proceed by induction on the number of leaves and the induction hypothesis applied is the one explained above the statement of the theorem. By the induction hypothesis we know that the rank of  $\mathbf{J}_{A|L_A}(T_A)$  is equal to  $\text{codim}(CV_{T_A})$ . By the claw  $d$ -tree hypothesis (7), we know that

$$\mathbf{J}_{L_B|B}(T_B) = \left[ \begin{array}{ccc} * & * & \\ * & * & \\ * & * & \end{array} \middle| \mathbf{J}_{L_B|B}^*(T_B) \right]$$

and the rank of  $\mathbf{J}_{L_B|B}^*(T_B)$  is equal to  $\text{codim}(CV_{T_B})$ . This assures that the first  $\text{codim}(CV_{T_B}) + \text{codim}(CV_{T_A})$  rows in the matrix are linearly independent. For the third block and by virtue of Lemma 4.8, we have

$$\left[ \frac{\partial}{\partial q_{i,j}^k} E_{i,j}^k \right] = \left[ \begin{array}{cccc} * & \dots & \dots & * \\ * & & & * \\ * & \dots & \dots & * \end{array} \middle| \text{Diag}(\Delta_k(\pi_n)) \right]$$

where  $\text{Diag}(\Delta_k(\pi_n))$  is the diagonal matrix with entries  $\{\Delta_k(\pi_n)\}_k$  and columns indexed by the coordinates  $q_{i,j}^k$  for  $i, j \geq m_k + 1$ . By Lemma 5.5 below, these entries are nonzero. We conclude that the rank of  $\mathbf{J}_{A|B}(T)$  is maximal and equal to  $\text{codim}(CV_T)$ .  $\square$

**Lemma 5.5** *Let  $\pi_n$  be a generic point of no evolution. Then,  $\Delta_k(\pi_n) \neq 0$  for every  $k = 1, \dots, t$ .*

PROOF. The matrix  $M_k^0$  for any tensor in  $p \in (\otimes^n W)^G$  represents a tensor in  $\mathcal{F}_{k^*}(W_\alpha) \otimes \mathcal{F}_k(W_\beta)$  obtained as follows:

1. first contract  $p$  with  $f_{L(T) \setminus \{\alpha, \beta\}}$  to obtain a tensor  $p'$  in  $(W_\alpha \otimes W_\beta)^G$ ,
2. project  $p'$  according to the decomposition  $(W_\alpha \otimes W_\beta)^G \cong \bigoplus_k \mathcal{F}_{k^*}(W_\alpha) \otimes \mathcal{F}_k(W_\beta)$ .

The determinant of  $M_k^0$  is nonzero if and only if the associated map  $\mathcal{F}_k(W_\alpha^*) \rightarrow \mathcal{F}_k(W_\beta)$  has maximal rank, i.e. is an isomorphism. However, this is the case for all  $k$  if and only if  $p'$  defines a  $G$ -isomorphism  $W_\alpha^* \rightarrow W_\beta$ .

By lemma 5.1, the marginalization of  $\pi_n = \sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} \mathbf{x} \otimes \dots \otimes \mathbf{x}$  over all leaves different from  $\alpha$  and  $\beta$  provides a tensor  $p' = \sum_{\mathbf{x} \in \Sigma} \pi_{\mathbf{x}} \mathbf{x} \otimes \mathbf{x}$ . This tensor corresponds to the map from  $W_\alpha^*$  to  $W_\beta$  whose matrix in basis  $\{X_i\}$  is diagonal with entries  $\pi_{X_i}$ . Therefore, if the coordinates  $\pi_X$  of  $\pi_n$  are all non-zero, this is an isomorphism and therefore the matrices  $M_k^0$  have non-zero determinant for all  $k$ . This proves the claim.  $\square$

**Remark 5.6** The list of equations for a phylogenetic tree as in Theorem 5.4 is obtained from edge invariants and from local equations for  $d$ -claw trees,  $d \in D$ . As pointed out in 4.7, the number of edge invariants is exponential in  $n$  (at least for the usual models) and can be substituted by a direct evaluation of the rank of the thin flattening. It is worth noticing that the other subset of equations, the ones coming from  $d$ -claw trees, are at most exponential in  $d$  and therefore this subset can reasonably be used in practice.

## 6 Explicit equations for usual models

The aim of this section is to provide explicit examples of complete intersections of the particular models listed in Example 3.4. For the Kimura 3-parameter, this was already done in [13]; here we deal with GMM, strand symmetric and Jukes-Cantor models (the only remaining case would be Kimura 2-parameter, for which the tripod assumption could be checked using computational algebra software).

As mentioned in Example 4.3, for these models we can use the Fourier basis to span the isotypic components of  $W$ . In this cases, we can identify  $\underline{\Sigma}$  with the group  $(\mathbb{Z}_2 \times \mathbb{Z}_2, +)$  via

$$\underline{A} \mapsto (0, 0); \quad \underline{C} \mapsto (1, 0); \quad \underline{G} \mapsto (0, 1); \quad \underline{T} \mapsto (1, 1). \quad (8)$$

We denote by  $\chi_A, \chi_C, \chi_G, \chi_T$  the characters associated to this group  $(\mathbb{Z}_2 \times \mathbb{Z}_2, +)$  (see table 2). These characters are useful to describe the coordinates of a point of no evolution.

	<u>A</u>	<u>C</u>	<u>G</u>	<u>T</u>
$\chi_A$	1	1	1	1
$\chi_C$	1	1	-1	-1
$\chi_G$	1	-1	1	-1
$\chi_T$	1	-1	-1	1

Table 2: Description of characters  $\chi_A, \chi_C, \chi_G, \chi_T$  for the group  $(\mathbb{Z}_2 \times \mathbb{Z}_2, +)$ .

**Lemma 6.1** *If  $\pi_n$  is a point of no evolution, then  $\pi_n = \sum_{\underline{Y}_1, \dots, \underline{Y}_n} q_{\underline{Y}_1, \dots, \underline{Y}_n} \underline{Y}_1 \dots \underline{Y}_n$ , where*

$$q_{\underline{Y}_1, \dots, \underline{Y}_n} = \frac{1}{4^n} (\pi_A + \chi_C(\underline{Y})\pi_C + \chi_G(\underline{Y})\pi_G + \chi_T(\underline{Y})\pi_T),$$

$\underline{Y} := \underline{Y}_1 + \dots + \underline{Y}_n$  with the operation given by the identification (8).

**PROOF.** From the definition of the Fourier basis in 4.1.1 and using Table 2, we have  $\mathbf{X} = \frac{1}{4} \sum_{\underline{Y} \in \underline{\Sigma}} \chi_{\mathbf{X}}(\underline{Y}) \underline{Y}$ , for any  $\mathbf{X} \in \Sigma$ . Now, using that characters of 1-dimensional representations are multiplicative, if  $\pi_n$  is a point of no evolution we have

$$\begin{aligned} \pi_n &= \sum_{\mathbf{X}} \pi_{\mathbf{X}} \mathbf{X} \otimes \dots \otimes \mathbf{X} = \\ &= \sum_{\mathbf{X}} \frac{\pi_{\mathbf{X}}}{4^n} \left( \sum_{\underline{Y}_1} \chi_{\mathbf{X}}(\underline{Y}_1) \underline{Y}_1 \right) \otimes \dots \otimes \left( \sum_{\underline{Y}_n} \chi_{\mathbf{X}}(\underline{Y}_n) \underline{Y}_n \right) = \\ &= \sum_{\mathbf{X}} \frac{\pi_{\mathbf{X}}}{4^n} \sum_{\underline{Y}_1, \underline{Y}_2, \dots, \underline{Y}_n} \chi_{\mathbf{X}}(\underline{Y}_1) \chi_{\mathbf{X}}(\underline{Y}_2) \dots \chi_{\mathbf{X}}(\underline{Y}_n) \underline{Y}_1 \underline{Y}_2 \dots \underline{Y}_n = \\ &= \sum_{\mathbf{X}} \frac{\pi_{\mathbf{X}}}{4^n} \sum_{\underline{Y}_1, \underline{Y}_2, \dots, \underline{Y}_n} \chi_{\mathbf{X}}(\underline{Y}_1 + \underline{Y}_2 + \dots + \underline{Y}_n) \underline{Y}_1 \underline{Y}_2 \dots \underline{Y}_n = \end{aligned}$$

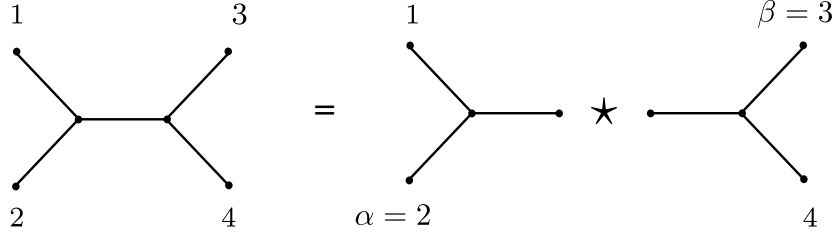


Figure 3: .

$$= \sum_{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n} \left( \frac{1}{4^n} \sum_{\mathbf{x}} \pi_{\mathbf{x}} \chi_{\mathbf{x}}(\mathbf{Y}) \right) \mathbf{Y}_1 \mathbf{Y}_2 \dots \mathbf{Y}_n.$$

From this, the claim follows.  $\square$

### 6.1 Equations for trees evolving under the Jukes-Cantor model

Take the Jukes-Cantor model, this is,  $G = \mathfrak{S}_4$ . There are five irreducible representations of the group  $G$ :  $N_1, \dots, N_5$  (see [31] §2.3). For each  $i = 1 \div 5$ , the representation  $N_i$  has the character  $\chi_i$  shown in Table 1. As in Example 4.3.5, we have  $u_1^1 = \underline{\mathbf{A}}$ ,  $u_1^4 = \underline{\mathbf{C}}$  and  $\mathcal{F}_1(W) = \langle \underline{\mathbf{A}} \rangle$  and  $\mathcal{F}_4(W) = \langle \underline{\mathbf{C}} \rangle$ . By letting the group  $G$  act, it follows that  $W = W[\chi_1] \oplus W[\chi_4]$ , with  $W[\chi_1] = \langle \underline{\mathbf{A}} \rangle$  and  $W[\chi_4] = \langle \underline{\mathbf{C}}, \underline{\mathbf{G}}, \underline{\mathbf{T}} \rangle$ .

We take the tree  $T$  with 4 leaves  $12|34$  and take the bipartition  $A = \{1, 2\}$  and  $B = \{3, 4\}$ . We proceed to obtain a basis for the ambient space  $(\otimes^4 W)^G$  of the variety  $V_T$ . Choose  $\alpha = 2$  and  $\beta = 3$  as in figure 3. Following the algorithm described, and noting that  $\chi^2 = 2\chi_1 + \chi_3 + 3\chi_4 + \chi_5$ , we obtain the following basis

$$\begin{array}{ll} u_{A,1}^1 = \underline{\mathbf{A}\mathbf{A}} & u_{B,1}^1 = \underline{\mathbf{A}\mathbf{A}} \\ u_{A,2}^1 = \underline{\mathbf{C}\mathbf{C}} + \underline{\mathbf{G}\mathbf{G}} + \underline{\mathbf{T}\mathbf{T}} & u_{B,2}^1 = \underline{\mathbf{C}\mathbf{C}} + \underline{\mathbf{G}\mathbf{G}} + \underline{\mathbf{T}\mathbf{T}} \\ u_{A,1}^3 = \underline{\mathbf{C}\mathbf{C}} - \underline{\mathbf{G}\mathbf{G}} & u_{B,1}^3 = \underline{\mathbf{C}\mathbf{C}} - \underline{\mathbf{G}\mathbf{G}} \\ u_{A,1}^4 = \underline{\mathbf{A}\mathbf{C}} & u_{B,1}^4 = \underline{\mathbf{C}\mathbf{A}} \\ u_{A,2}^4 = \underline{\mathbf{C}\mathbf{A}} & u_{B,2}^4 = \underline{\mathbf{A}\mathbf{C}} \\ u_{A,3}^4 = \underline{\mathbf{G}\mathbf{T}} + \underline{\mathbf{T}\mathbf{G}} & u_{B,3}^4 = \underline{\mathbf{G}\mathbf{T}} + \underline{\mathbf{T}\mathbf{G}} \\ u_{A,1}^5 = \underline{\mathbf{G}\mathbf{T}} - \underline{\mathbf{T}\mathbf{G}} & u_{B,1}^5 = \underline{\mathbf{G}\mathbf{T}} - \underline{\mathbf{T}\mathbf{G}} \end{array}$$

Now, for each of the irreducible representations of  $\mathfrak{S}_4$ , we can choose a subgroup  $H_k$  so that  $\{hu_{A,i}^k \mid h \in H_k\}$  is a basis of the representation spanned by the  $\mathfrak{S}_4$ -orbit of  $u_{A,i}^k$ . Namely,

$$\begin{array}{ll} H_1 = \{e\} & n_1 = 1; \\ H_3 = \{e, (\mathbf{AC})\} & n_3 = 2; \\ H_4 = \{e, (\mathbf{CGT}), (\mathbf{CTG})\} & n_4 = 3; \\ H_5 = \{e, (\mathbf{CGT}), (\mathbf{CTG})\} & n_5 = 3. \end{array}$$

A basis for  $(W_A \otimes W_B)^G$  is inferred from  $\bigoplus_k \mathcal{F}_k(W_A) \otimes \mathcal{F}_k(W_B)$  by taking the image of the operator  $S$  specified in Remark 4.1 applied to the tensors  $u_{A,i}^k \otimes u_{B,j}^k$ :

$$S(u_{A,i}^k \otimes u_{B,j}^k) := \frac{n_k}{|G|} \sum_{g \in G} (g \cdot u_{A,i}^k) \otimes (g \cdot u_{B,j}^k),$$

that is,

$$\begin{aligned}
S(u_{A,1}^1 \otimes u_{B,1}^1) &= \underline{\text{AAAA}}, \\
S(u_{A,1}^1 \otimes u_{B,2}^1) &= \frac{1}{3} \underline{\text{AA}} \otimes (\underline{\text{CC}} + \underline{\text{GG}} + \underline{\text{TT}}), \\
S(u_{A,2}^1 \otimes u_{B,1}^1) &= \frac{1}{3} (\underline{\text{CC}} + \underline{\text{GG}} + \underline{\text{TT}}) \otimes \underline{\text{AA}}, \\
S(u_{A,2}^1 \otimes u_{B,2}^1) &= \frac{1}{3} (\underline{\text{CC}} + \underline{\text{GG}} + \underline{\text{TT}}) \otimes (\underline{\text{CC}} + \underline{\text{GG}} + \underline{\text{TT}}), \\
S(u_{A,1}^3 \otimes u_{B,1}^3) &= \frac{2}{3} ((\underline{\text{CC}} - \underline{\text{GG}}) \otimes (\underline{\text{CC}} - \underline{\text{GG}}) + (\underline{\text{GG}} - \underline{\text{TT}}) \otimes (\underline{\text{GG}} - \underline{\text{TT}}) + (\underline{\text{TT}} - \underline{\text{CC}}) \otimes (\underline{\text{TT}} - \underline{\text{CC}})) \\
S(u_{A,1}^4 \otimes u_{B,1}^4) &= \underline{\text{ACCA}} + \underline{\text{AGGA}} + \underline{\text{ATTA}}, \\
S(u_{A,1}^4 \otimes u_{B,2}^4) &= \underline{\text{ACAC}} + \underline{\text{AGAG}} + \underline{\text{ATAT}}, \\
S(u_{A,1}^4 \otimes u_{B,3}^4) &= \underline{\text{AC}} \otimes (\underline{\text{GT}} + \underline{\text{TG}}) + \underline{\text{AG}} \otimes (\underline{\text{CT}} + \underline{\text{TC}}) + \underline{\text{AT}} \otimes (\underline{\text{CG}} + \underline{\text{GC}}), \\
S(u_{A,2}^4 \otimes u_{B,1}^4) &= \underline{\text{CACA}} + \underline{\text{GAGA}} + \underline{\text{TATA}}, \\
S(u_{A,2}^4 \otimes u_{B,2}^4) &= \underline{\text{CAAC}} + \underline{\text{GAAG}} + \underline{\text{TAAT}}, \\
S(u_{A,2}^4 \otimes u_{B,3}^4) &= \underline{\text{CA}} \otimes (\underline{\text{GT}} + \underline{\text{TG}}) + \underline{\text{GA}} \otimes (\underline{\text{CT}} + \underline{\text{TC}}) + \underline{\text{TA}} \otimes (\underline{\text{CG}} + \underline{\text{GC}}), \\
S(u_{A,4}^4 \otimes u_{B,1}^4) &= (\underline{\text{GT}} + \underline{\text{TG}}) \otimes \underline{\text{CA}} + (\underline{\text{CT}} + \underline{\text{TC}}) \otimes \underline{\text{GA}} + (\underline{\text{CG}} + \underline{\text{GC}}) \otimes \underline{\text{TA}}, \\
S(u_{A,3}^4 \otimes u_{B,2}^4) &= (\underline{\text{GT}} + \underline{\text{TG}}) \otimes \underline{\text{AC}} + (\underline{\text{CT}} + \underline{\text{TC}}) \otimes \underline{\text{AG}} + (\underline{\text{CG}} + \underline{\text{GC}}) \otimes \underline{\text{AT}}, \\
S(u_{A,3}^4 \otimes u_{B,3}^4) &= (\underline{\text{GT}} + \underline{\text{TG}}) \otimes (\underline{\text{GT}} + \underline{\text{TG}}) + (\underline{\text{CT}} + \underline{\text{TC}}) \otimes (\underline{\text{CT}} + \underline{\text{TC}}) + (\underline{\text{CG}} + \underline{\text{GC}}) \otimes (\underline{\text{CG}} + \underline{\text{GC}}), \\
S(u_{A,1}^5 \otimes u_{B,1}^5) &= (\underline{\text{GT}} - \underline{\text{TG}}) \otimes (\underline{\text{GT}} - \underline{\text{TG}}) + (\underline{\text{CT}} - \underline{\text{TC}}) \otimes (\underline{\text{CT}} - \underline{\text{TC}}) + (\underline{\text{CG}} - \underline{\text{GC}}) \otimes (\underline{\text{CG}} - \underline{\text{GC}}).
\end{aligned}$$

As above, denote by  $q_{i,j}^k$  the coordinates corresponding to this basis  $S(u_{A,i}^k \otimes u_{B,j}^k)$ . We proceed to obtain a complete intersection for the tree  $T$  with 4 leaves 12|34. Take the bipartition  $A = \{1, 2\}$  and  $B = \{3, 4\}$ .

First of all, we proceed to obtain the edge invariants following the section 4.2. If  $\pi_4$  is a no evolution point, write  $\pi_4 = \sum_{k,i,j} q_{ij}^k S(u_{A,i}^k \otimes u_{B,j}^k)$ . Each irreducible representation  $N_k$  of  $\mathfrak{S}_4$  gives rise to a  $m_k(2) \times m_k(2)$ -matrix  $M_k$ :

$$M_1 = \begin{pmatrix} q_{11}^1 & q_{12}^1 \\ q_{21}^1 & q_{22}^1 \end{pmatrix}, \quad M_3 = \begin{pmatrix} q_{11}^3 \end{pmatrix}, \quad M_4 = \begin{pmatrix} q_{11}^4 & q_{12}^4 & q_{13}^4 \\ q_{21}^4 & q_{22}^4 & q_{23}^4 \\ q_{31}^4 & q_{32}^4 & q_{33}^4 \end{pmatrix}, \quad M_5 = \begin{pmatrix} q_{11}^5 \end{pmatrix}$$

where the rows of each  $M_k$  are indexed by the  $\{u_{A,i}^k\}$  and the columns are indexed by the  $\{u_{B,j}^k\}$ . Notice that there is no  $M_2$  as there is no isotypic component corresponding to  $N_2$  in  $\otimes^4 W$ . Moreover, from Lemma 5.5, we know that  $\Delta_1(\pi_4) = q_{11}^1 \neq 0$  and  $\Delta_4(\pi_4) = q_{11}^4 \neq 0$ . The resulting edge invariants arise as rank restrictions for these matrices:

$$\begin{aligned}
\chi_1 : q_{2,2}^1 q_{1,1}^1 - q_{1,2}^1 q_{2,1}^1 &= 0 & \chi_4 : q_{2,2}^4 q_{1,1}^4 - q_{1,2}^4 q_{2,1}^4 &= 0 \\
\chi_3 : q_{1,1}^3 &= 0 & q_{2,3}^4 q_{1,1}^4 - q_{1,3}^4 q_{2,1}^4 &= 0 \\
\chi_5 : q_{1,1}^5 &= 0 & q_{3,2}^4 q_{1,1}^4 - q_{1,2}^4 q_{3,1}^4 &= 0 \\
& & q_{3,3}^4 q_{1,1}^4 - q_{1,3}^4 q_{3,1}^4 &= 0
\end{aligned}$$

We also need to consider the equations obtained from the tripods associated to the bipartition:  $T_A$  with leaves  $\{1, \alpha, L_A\}$ , and  $T_B$  with leaves  $\{L_B, \beta, 4\}$  (see figure 3). Now, it can be seen that a complete intersection for the tripod  $T_3$  with leaves  $x, y, z$  is given by

$$Q_{1,1}^4 Q_{1,2}^4 Q_{1,2}^1 - Q_{1,1}^1 (Q_{1,3}^4)^2 = 0$$

where these  $Q_{ij}^k$  are the coordinates corresponding to the basis linked to the edge split  $x|yz$ :

$$\begin{aligned}
S(u_{x,1}^1 \otimes u_{yz,1}^1) &= \underline{AAA}, \\
S(u_{x,1}^1 \otimes u_{yz,2}^1) &= \underline{ACC} + \underline{AGG} + \underline{ATT}, \\
S(u_{x,1}^4 \otimes u_{yz,1}^4) &= \underline{CAC} + \underline{GAG} + \underline{TAT}, \\
S(u_{x,1}^4 \otimes u_{yz,2}^4) &= \underline{CCA} + \underline{GGA} + \underline{TTA}, \\
S(u_{x,1}^4 \otimes u_{yz,3}^4) &= \underline{CGT} + \underline{CTG} + \underline{GCT} + \underline{GTC} + \underline{TCG} + \underline{TGC}.
\end{aligned}$$

From  $V_{T_A}$ , we take  $x = L_A, y = 1, z = 2$  to obtain the extended equation in the original coordinates  $q_{ij}^k$ :

$$q_{1,1}^4 q_{2,1}^4 q_{1,2}^1 - q_{1,1}^1 (q_{3,1}^4)^2 = 0.$$

Analogously, from  $V_{T_B}$ , we take  $x = L_B, y = 3, z = 4$  to obtain the extended equation:

$$q_{1,1}^4 q_{1,2}^4 q_{1,2}^1 - q_{1,1}^1 (q_{1,3}^4)^2 = 0.$$

These 9 equations define a local complete intersection around generic points of no evolution for the variety  $CV_G(T)$  in  $(\otimes^4 W)^G$ .

## 6.2 Equations for the tripod evolving under the general Markov model

The aim of this subsection is to explicitly provide codimension many equations of the variety  $X$  for the tripod for GMM that cut  $X$  out in a neighborhood of no evolution points. As the salmon conjecture is well-studied and answered on the set theoretic level [37, 11, 29, 45] many of the equations of  $X$  are known. However, it turns out that the simplest equations, going back to Strassen, are enough to obtain a description of the local complete intersection. As we will see not only are they enough - also their choice is astonishingly natural.

Recall that when  $\Sigma$  has  $\kappa$  elements,  $X$  is the  $\kappa$ -th secant variety of the Segre product  $\mathbb{P}(A) \times \mathbb{P}(B) \times \mathbb{P}(C)$ , i.e. the closure of the locus of rank  $\kappa$  tensors in the space  $\mathbb{P}(A \otimes B \otimes C)$ , where  $\dim A = \dim B = \dim C = \kappa$ .

For the sake of completeness let us recall Landsberg-Ottaviani's interpretation of Strassen equations [38]. Each tensor  $\tau \in A \otimes B \otimes C$  is naturally identified with a map  $\tau : A^* \rightarrow B \otimes C$ . Let us tensor this map with the identity on  $C$  obtaining a map  $A^* \otimes C \rightarrow B \otimes C \otimes C$ . Using the natural map  $C \otimes C \rightarrow C \wedge C$  we obtain  $f(\tau) : A^* \otimes C \rightarrow B \otimes (C \wedge C)$ .

When  $\tau$  has rank one, i.e.  $\tau = a \otimes b \otimes c$ , then the rank of  $f(\tau)$  (as a matrix) is at most  $\kappa - 1$  because the image of  $f(\tau)$  is the subspace  $b \otimes (c \wedge C)$ . Hence, if  $\tau$  is of rank  $\kappa$ , then  $f(\tau)$  has rank at most  $\kappa(\kappa - 1)$ . Using the matrix representation of  $f(\tau)$  all  $\kappa(\kappa - 1) + 1$  minors provide equations of the  $\kappa$ -th secant variety.

In coordinates, if  $\tau = \sum a_{ijk} X_i \otimes X_j \otimes X_k$ , in a certain basis  $X_1, \dots, X_\kappa$  of  $W$ , then the entry of the column  $X_i^* \otimes X_j$  and row  $X_s \otimes (X_r \wedge X_t)$  equals (we assume  $r < t$ ):

- 0, if  $r$  and  $t$  are different from  $j$ ;
- $-a_{ist}$ , if  $r = j$ ;
- $a_{isr}$ , if  $t = j$ ;

because  $f(\tau)(X_i^* \otimes X_j) = \sum_{p,q} a_{ipq} X_p \otimes (X_q \wedge X_j)$ . A display of this matrix is shown in the Table 3.



**Definition 6.2** In the matrix representation of  $f(\tau)$  exactly  $\kappa(\kappa - 1)$  columns contain an entry  $a_{iii}$  or  $-a_{iii}$  for some  $i$ . Namely, these are the columns indexed by  $X_i^* \otimes X_j$ , where  $i \neq j$ . Each such column contains exactly one such entry. Moreover, these entries are contained in  $\kappa(\kappa - 1)$  different rows: those indexed by  $X_i \otimes (X_j \wedge X_i)$  (for  $j < i$ ) or  $X_i \otimes (X_i \wedge X_j)$  (for  $j > i$ ). These precise rows and columns will be called *distinguished*. In table 3 distinguished rows and columns are marked with  $*$  and depicted in gray.

Consider any minor  $M$  of  $f(\tau)$  of order  $\kappa(\kappa - 1) + 1$ . If it does not contain all the distinguished rows and columns, then all its derivatives vanish on any point of no evolution. Indeed, each monomial in  $M$  will contain at least a degree two factor in variables different from  $a_{iii}$ , hence any derivative of such monomial (if nonzero) will contain such a variable and will vanish on the no evolution point.

Thus from now on we will be interested only in those minors  $M$  that contain all the distinguished rows and columns. Such minors are of course specified by choosing a non-distinguished row  $r$  and column  $c$ . By the same argument as above, only the derivative of  $M$  with respect to the  $(r, c)$  entry can be nonzero at a point of no evolution: this derivative equals the determinant of the submatrix given by distinguished rows and columns, i.e. it equals

$$c := \pm \left( \prod_i a_{iii} \right)^{\kappa-1}. \quad (9)$$

Let us consider a nondistinguished row indexed by  $X_s \otimes (X_r \wedge X_t)$ ,  $r < t$ ,  $r \neq s$ ,  $t \neq s$ . It contains exactly two nonzero variables in the nondistinguished columns:  $-a_{rst}$  and  $a_{tsr}$ . In particular, there are  $2(\kappa \binom{\kappa}{2} - \kappa(\kappa - 1)) = \kappa(\kappa - 1)(\kappa - 2)$  variables in the submatrix indexed by nondistinguished rows and columns, and they are all different. Hence, the corresponding minors have independent differentials at a generic point of no evolution (which does not have any coordinate equal to zero). Notice however that  $\kappa(\kappa - 1)(\kappa - 2) = \kappa^3 - 1 - (3\kappa(\kappa - 1) + \kappa - 1)$  is the codimension of the variety by Terracini's lemma [20, 1]. Thus we can conclude that the given minors provide locally a description of the variety as a complete intersection at the generic points of no evolution.

For  $r, s, t$  in  $\{1, \dots, \kappa\}$ , with  $s \neq r$ ,  $t \neq r, s$ , we call  $\text{eq}_{X_r, X_s, X_t}$  the equation given by the minor formed by the distinguished rows and columns, plus row  $X_s \otimes (X_r \wedge X_t)$ , and column  $X_r^* \otimes X_r$  if  $r < t$  or  $X_t^* \otimes X_t$  if  $t < r$ . We have proven the following:

**Lemma 6.3** *The equations  $\text{eq}_{X_r, X_s, X_t}$  for  $s \neq r$ ,  $t \neq r, s$  describe the variety  $CV_{GMM}(T_3)$  as a complete intersection locally at a generic point of no evolution.*

Next, we proceed to prove that Assumption 5.2 is also satisfied. Consider  $\kappa = 4$  and let  $\{\underline{A}, \underline{C}, \underline{G}, \underline{T}\}$  be the Fourier basis. We have 24 equations  $\text{eq}_{\underline{X}, \underline{Y}, \underline{Z}}$  which are indexed by 3-element subsets of  $\{\underline{A}, \underline{C}, \underline{G}, \underline{T}\}$ . By the previous discussion, each equation  $\text{eq}_{\underline{X}, \underline{Y}, \underline{Z}}$  has only one nonzero directional derivative (after evaluated at a generic point of no evolution) in the standard basis, namely with respect to  $\underline{X} \otimes \underline{Y} \otimes \underline{Z}$ . Notice that if we removed all the columns of the Jacobian matrix indexed by variables of type  $\underline{A} \otimes \underline{Y} \otimes \underline{Z}$  the matrix would drop rank: all rows indexed by equations  $\text{eq}_{\underline{A}, \underline{Y}, \underline{Z}}$  would be zero.

Let us consider the basis  $\underline{X} \otimes \underline{Y} \otimes \underline{Z}$ , where  $\underline{X} \in \{\underline{A}, \underline{C}, \underline{G}, \underline{T}\}$ , but  $\underline{Y}, \underline{Z} \in \{\underline{A}, \underline{C}, \underline{G}, \underline{T}\}$ . Let us call  $J$  the Jacobian matrix with 24 rows indexed by the above equations written in this new basis and 64 columns indexed by basis elements  $\underline{X} \otimes \underline{Y} \otimes \underline{Z}$ . Let  $\tilde{J}$  be the submatrix of  $J$  obtained by removing all the columns indexed by a variable of type  $\underline{A} \otimes \underline{Y} \otimes \underline{Z}$  for any  $\underline{Y}, \underline{Z}$ .

**Lemma 6.4** *The rank of the matrix  $\tilde{J}$  is maximal, i.e. equal to 24.*

PROOF. Let us fix distinct  $Y, Z \in \{A, C, G, T\}$ . There are precisely two equations  $eq_{X_1, Y, Z}, eq_{X_2, Y, Z}$  as described above. The evaluation at a generic point of no evolution of the directional derivatives of these equations with respect to  $\underline{S}VW$  equals zero unless  $V = Y$  and  $W = Z$ . These give us 3 variables that can give nonzero derivatives (as we assume  $\underline{S} \neq \underline{A}$ ). Actually, the directional derivative of  $eq_{XYZ}$  with respect to  $\underline{S} \otimes V \otimes W$  is equal to

$$\frac{deq_{XYZ}}{d\underline{S} \otimes V \otimes W} = \begin{cases} \varepsilon c & \text{if } V = Y, W = Z \\ 0 & \text{otherwise.} \end{cases}$$

where  $c$  is the amount defined in (9) and  $\varepsilon$  represents the sign of  $X$  in  $\underline{S}$ :  $\varepsilon = 1$  if  $X = A$  or  $X = S$ , and  $\varepsilon = -1$  otherwise.

On the other hand, at a generic point of no evolution, the evaluation of the derivatives with respect to any variable  $\underline{S} \otimes Y \otimes Z$  gives a nonzero value only when applied to the equations  $eq_{X_1, Y, Z}$  or  $eq_{X_2, Y, Z}$ . Hence the nonzero entries of the  $24 \times 48$  matrix  $\tilde{J}$  are contained in 12 rectangles of shape  $2 \times 3$ , not sharing rows or columns. To finish the proof it remains to show a  $2 \times 2$  submatrix with nonzero determinant in each rectangle.

If  $Y = A$  or  $Z = A$ , then  $X_1, X_2 \neq A$ . Without loss of generality, we may assume that  $Y = A$ . Then, we choose the derivatives with respect to  $\underline{X}_1YZ$ , and  $\underline{X}_2YZ$ . The submatrix obtained has the form:

$$\begin{pmatrix} c & -c \\ -c & -c \end{pmatrix}.$$

If  $Y, Z \neq A$ , then  $A \in \{X_1, X_2\}$  and we can take  $X_1 = A$ . We choose the derivatives with respect to  $\underline{X}_2YZ$  and  $\underline{Y}YZ$ . The obtained submatrix is of the form:

$$\begin{pmatrix} c & c \\ c & -c \end{pmatrix}.$$

In any case, the determinant is  $-2c^2 \neq 0$  and we are done.  $\square$

### 6.3 Equations for trees evolving under the strand symmetric model

Take  $G = \langle (AT)(CG) \rangle \cong \mathbb{Z}_2$ , corresponding to the strand symmetric model as in Example 3.4.

In order to deduce equations for the tripod, we construct first a convenient basis for the space  $(\otimes^3 W)^G$ . We take  $\mathcal{F}_1(W_\alpha) = \langle \underline{A}, \underline{T} \rangle$  and  $\mathcal{F}_2(W_\alpha) = \langle \underline{C}, \underline{G} \rangle$ . Take  $A = \{1, 2\}$  and keep the notation used in Example 4.5. A basis for  $(W_A \otimes W)^G$  would be inferred from  $\bigoplus_k \mathcal{F}_k(W_A) \otimes \mathcal{F}_k(W)$  by taking the image of the operator  $S$  specified in Remark 4.1 applied to the tensors  $u_{A,i}^k \otimes u_j^k$  (as the irreducible representations have dimension 1, Remark 4.1 trivially applies):

$S(u_{A,1}^1 \otimes u_1^1) = \underline{AAA}$	$S(u_{A,5}^1 \otimes u_1^1) = \underline{CCA}$	$S(u_{A,1}^1 \otimes u_2^1) = \underline{AAT}$	$S(u_{A,5}^1 \otimes u_2^1) = \underline{CCT}$
$S(u_{A,2}^1 \otimes u_1^1) = \underline{ATA}$	$S(u_{A,6}^1 \otimes u_1^1) = \underline{CGA}$	$S(u_{A,2}^1 \otimes u_2^1) = \underline{ATT}$	$S(u_{A,6}^1 \otimes u_2^1) = \underline{CGT}$
$S(u_{A,3}^1 \otimes u_1^1) = \underline{TAA}$	$S(u_{A,7}^1 \otimes u_1^1) = \underline{GCA}$	$S(u_{A,3}^1 \otimes u_2^1) = \underline{TAT}$	$S(u_{A,7}^1 \otimes u_2^1) = \underline{GCT}$
$S(u_{A,4}^1 \otimes u_1^1) = \underline{TTA}$	$S(u_{A,8}^1 \otimes u_1^1) = \underline{GGA}$	$S(u_{A,4}^1 \otimes u_2^1) = \underline{TTT}$	$S(u_{A,8}^1 \otimes u_2^1) = \underline{GGT}$
$S(u_{A,1}^2 \otimes u_1^2) = \underline{ACC}$	$S(u_{A,5}^2 \otimes u_1^2) = \underline{CAC}$	$S(u_{A,1}^2 \otimes u_2^2) = \underline{ACG}$	$S(u_{A,5}^2 \otimes u_2^2) = \underline{CAG}$
$S(u_{A,2}^2 \otimes u_1^2) = \underline{AGC}$	$S(u_{A,6}^2 \otimes u_1^2) = \underline{CTC}$	$S(u_{A,2}^2 \otimes u_2^2) = \underline{AGG}$	$S(u_{A,6}^2 \otimes u_2^2) = \underline{CTG}$
$S(u_{A,3}^2 \otimes u_1^2) = \underline{TCC}$	$S(u_{A,7}^2 \otimes u_1^2) = \underline{GAC}$	$S(u_{A,3}^2 \otimes u_2^2) = \underline{TCG}$	$S(u_{A,7}^2 \otimes u_2^2) = \underline{GAG}$
$S(u_{A,4}^2 \otimes u_1^2) = \underline{TGC}$	$S(u_{A,8}^2 \otimes u_1^2) = \underline{GTC}$	$S(u_{A,4}^2 \otimes u_2^2) = \underline{TGG}$	$S(u_{A,8}^2 \otimes u_2^2) = \underline{GTG}$

In other words, the *Fourier basis for SSM* is the subbasis of the usual Fourier basis for  $\otimes^3 W$  formed by triplets that contain an even number of elements in  $\{\underline{C}, \underline{G}\}$ . Thus we denote its coordinates as

the usual Fourier coordinates  $q_{\underline{XYZ}}$  (corresponding to the basis vector  $\underline{XYZ}$ ). If  $\pi_3 = \sum_X \pi_X \underline{XXX}$  is a no evolution point, then its Fourier coordinates are given by

$$q_{\underline{XYZ}} = \begin{cases} 2\pi^+, & \text{if } \underline{X} + \underline{Y} + \underline{Z} = \underline{A} \\ 2\pi^-, & \text{if } \underline{X} + \underline{Y} + \underline{Z} = \underline{T} \end{cases}$$

where  $\pi^+ := \pi_A + \pi_C$  and  $\pi^- := \pi_A - \pi_C$  (see Lemma 6.1) and the sum of nucleotides is done according to (8).

A complete intersection for the variety corresponding to the tripod (which has codimension 12 in  $(\otimes^3 W)^G$ ) is defined by 12 equations and can be obtained from the 24 equations we described for the tripod evolving under GMM as follows. As in Section 6.2 we consider the tensor  $T$  and write the matrix  $f(T)$  in the Fourier coordinates. As  $q_{\underline{XYZ}}$  is 0 if  $\underline{XYZ}$  contains an odd number of elements in  $\{\underline{C}, \underline{G}\}$ , the matrix  $f(T)$  reduces to the matrix presented in Table 4. The subindices 1, 2, 3, 4 refer now to  $\underline{A}, \underline{C}, \underline{G}, \underline{T}$  respectively. When we consider the same equations as in the previous section we observe that out of the 12 nondistinguished rows, only 6 contain nonzero entries at the nondistinguished columns (and they contain exactly 2 nonzero entries). These rows are those labelled by  $\underline{X}_r \otimes (\underline{X}_s \wedge \underline{X}_t)$  such that  $s < t$  and  $\{\underline{X}_r, \underline{X}_s, \underline{X}_t\}$  contains an even number of  $\underline{C}, \underline{G}$ 's. The same argument used for the general Markov model proves that these twelve  $13 \times 13$  minors define a local complete intersection at the generic points of evolution and that they satisfy assumption 5.2.

Now out of these equations we provide a local complete intersection for trees with four leaves. Take  $T$  the tree with 4 leaves and choose  $\alpha = 2$  and  $\beta = 3$  as in the previous example (see figure 3). We proceed to obtain a basis for the ambient space  $(\otimes^4 W)^G$  of the variety  $V_T$ . Similar computations as above show that

$$\begin{aligned} \mathcal{F}_1(W_B) &= \langle \underline{AA}, \underline{TA}, \underline{AT}, \underline{TT}, \underline{CC}, \underline{CG}, \underline{GC}, \underline{GG} \rangle; \\ \mathcal{F}_2(W_B) &= \langle \underline{CA}, \underline{GA}, \underline{CT}, \underline{GT}, \underline{AC}, \underline{TC}, \underline{AG}, \underline{TG} \rangle. \end{aligned}$$

So, a basis for  $(W_A \otimes W_B)^G$  is given by all tensors of the form  $\omega_{i,j}^k := u_i^k \otimes u_j^k$  (as the irreducible representations have dimension 1, Remark 4.1 trivially applies).

Now, if  $\pi_4 \in (\otimes^4 W)^G$  is a no evolution point, and we write  $\pi_4 = \sum_{k,i,j} q_{ij}^k \omega_{ij}^k$ , then each irreducible representation gives rise to a  $m_k(2) \times m_k(2)$ -matrix  $M_k$ ,  $k = 1, 2$ :

$$M_k = \begin{pmatrix} q_{11}^k & q_{12}^k & q_{13}^k & q_{14}^k & q_{15}^k & q_{16}^k & q_{17}^k & q_{18}^k \\ q_{21}^k & q_{22}^k & q_{23}^k & q_{24}^k & q_{25}^k & q_{26}^k & q_{27}^k & q_{28}^k \\ q_{31}^k & q_{32}^k & q_{33}^k & q_{34}^k & q_{35}^k & q_{36}^k & q_{37}^k & q_{38}^k \\ q_{41}^k & q_{42}^k & q_{43}^k & q_{44}^k & q_{45}^k & q_{46}^k & q_{47}^k & q_{48}^k \\ q_{51}^k & q_{52}^k & q_{53}^k & q_{54}^k & q_{55}^k & q_{56}^k & q_{57}^k & q_{58}^k \\ q_{61}^k & q_{62}^k & q_{63}^k & q_{64}^k & q_{65}^k & q_{66}^k & q_{67}^k & q_{68}^k \\ q_{71}^k & q_{72}^k & q_{73}^k & q_{74}^k & q_{75}^k & q_{76}^k & q_{77}^k & q_{78}^k \\ q_{81}^k & q_{82}^k & q_{83}^k & q_{84}^k & q_{85}^k & q_{86}^k & q_{87}^k & q_{88}^k \end{pmatrix}$$

where rows are indexed by the  $\{u_{A,i}^k\}$  and columns are indexed by the  $\{u_{B,j}^k\}$ . As above, from Lemma 5.5, we know that

$$\Delta_1(\pi_4) = \begin{vmatrix} q_{11}^1 & q_{12}^1 \\ q_{21}^1 & q_{22}^1 \end{vmatrix} \neq 0, \quad \text{and} \quad \Delta_2(\pi_4) = \begin{vmatrix} q_{11}^2 & q_{12}^2 \\ q_{21}^2 & q_{22}^2 \end{vmatrix} \neq 0.$$

The resulting edge invariants arise from each  $M_k$  as rank restrictions for the 3-minors containing  $\Delta_k(\pi_4)$ , namely:

$$q_{ij}^k \Delta_k(\pi_4) - q_{2j}^k (q_{11}^k q_{i2}^k - q_{12}^k q_{i1}^k) + q_{1j}^k (q_{21}^k q_{i2}^k - q_{22}^k q_{i1}^k) = 0, \quad \text{for } i, j \geq 3, \quad k = 1, 2.$$

These invariants together with the 12 equations obtained from  $T_A$  and the 12 equations obtained from  $T_B$  define a complete intersection for the variety of  $T$  locally near the points of no evolution.

	$A^* \otimes A$	$(*)$ $A^* \otimes C$	$(*)$ $A^* \otimes G$	$(*)$ $A^* \otimes T$	$(*)$ $C^* \otimes A$	$C^* \otimes C$	$(*)$ $C^* \otimes G$	$(*)$ $C^* \otimes T$	$(*)$ $G^* \otimes A$	$(*)$ $G^* \otimes C$	$G^* \otimes G$	$(*)$ $G^* \otimes T$	$(*)$ $T^* \otimes A$	$(*)$ $T^* \otimes C$	$(*)$ $T^* \otimes G$	$T^* \otimes T$
$A \otimes (A \wedge C)(*)$	$-a_{112}$	$a_{111}$	0	0	$-a_{212}$	$a_{211}$	0	0	$-a_{312}$	$a_{311}$	0	0	$-a_{412}$	$a_{411}$	0	0
$A \otimes (A \wedge G)(*)$	$-a_{113}$	0	$a_{111}$	0	$-a_{213}$	0	$a_{211}$	0	$-a_{313}$	0	$a_{311}$	0	$-a_{413}$	0	$a_{411}$	0
$A \otimes (A \wedge T)(*)$	$-a_{114}$	0	0	$a_{111}$	$-a_{214}$	0	0	$a_{211}$	$-a_{314}$	0	0	$a_{311}$	$-a_{414}$	0	0	$a_{411}$
$A \otimes (C \wedge G)$	0	$-a_{113}$	$a_{112}$	0	0	$-a_{213}$	$a_{212}$	0	0	$-a_{313}$	$a_{312}$	0	0	$-a_{413}$	$a_{412}$	0
$A \otimes (C \wedge T)$	0	$-a_{114}$	0	$a_{112}$	0	$-a_{214}$	0	$a_{212}$	0	$-a_{314}$	0	$a_{312}$	0	$-a_{414}$	0	$a_{412}$
$A \otimes (G \wedge T)$	0	0	$-a_{114}$	$a_{113}$	0	0	$-a_{214}$	$a_{213}$	0	0	$-a_{314}$	$a_{313}$	0	0	$-a_{414}$	$a_{413}$
$C \otimes (A \wedge C)(*)$	$-a_{122}$	$a_{121}$	0	0	$-a_{222}$	$a_{221}$	0	0	$-a_{322}$	$a_{321}$	0	0	$-a_{422}$	$a_{421}$	0	0
$C \otimes (A \wedge G)$	$-a_{123}$	0	$a_{121}$	0	$-a_{223}$	0	$a_{221}$	0	$-a_{323}$	0	$a_{321}$	0	$-a_{423}$	0	$a_{421}$	0
$C \otimes (A \wedge T)$	$-a_{124}$	0	0	$a_{121}$	$-a_{224}$	0	0	$a_{221}$	$-a_{324}$	0	0	$a_{321}$	$-a_{424}$	0	0	$a_{421}$
$C \otimes (C \wedge G)(*)$	0	$-a_{123}$	$a_{122}$	0	0	$-a_{223}$	$a_{222}$	0	0	$-a_{323}$	$a_{322}$	0	0	$-a_{423}$	$a_{422}$	0
$C \otimes (C \wedge T)(*)$	0	$-a_{124}$	0	$a_{122}$	0	$-a_{224}$	0	$a_{222}$	0	$-a_{324}$	0	$a_{322}$	0	$-a_{424}$	0	$a_{422}$
$C \otimes (G \wedge T)$	0	0	$-a_{124}$	$a_{123}$	0	0	$-a_{224}$	$a_{223}$	0	0	$-a_{324}$	$a_{323}$	0	0	$-a_{424}$	$a_{423}$
$G \otimes (A \wedge C)$	$-a_{132}$	$a_{131}$	0	0	$-a_{232}$	$a_{231}$	0	0	$-a_{332}$	$a_{331}$	0	0	$-a_{432}$	$a_{431}$	0	0
$G \otimes (A \wedge G)(*)$	$-a_{133}$	0	$a_{131}$	0	$-a_{233}$	0	$a_{231}$	0	$-a_{333}$	0	$a_{331}$	0	$-a_{433}$	0	$a_{431}$	0
$G \otimes (A \wedge T)$	$-a_{134}$	0	0	$a_{131}$	$-a_{234}$	0	0	$a_{231}$	$-a_{334}$	0	0	$a_{331}$	$-a_{434}$	0	0	$a_{431}$
$G \otimes (C \wedge G)(*)$	0	$-a_{133}$	$a_{132}$	0	0	$-a_{233}$	$a_{232}$	0	0	$-a_{333}$	$a_{332}$	0	0	$-a_{433}$	$a_{432}$	0
$G \otimes (C \wedge T)$	0	$-a_{134}$	0	$a_{132}$	0	$-a_{234}$	0	$a_{232}$	0	$-a_{334}$	0	$a_{332}$	0	$-a_{434}$	0	$a_{432}$
$G \otimes (G \wedge T)(*)$	0	0	$-a_{134}$	$a_{133}$	0	0	$-a_{234}$	$a_{233}$	0	0	$-a_{334}$	$a_{333}$	0	0	$-a_{434}$	$a_{433}$
$T \otimes (A \wedge C)$	$-a_{142}$	$a_{141}$	0	0	$-a_{242}$	$a_{241}$	0	0	$-a_{342}$	$a_{341}$	0	0	$-a_{442}$	$a_{441}$	0	0
$T \otimes (A \wedge G)$	$-a_{143}$	0	$a_{141}$	0	$-a_{243}$	0	$a_{241}$	0	$-a_{343}$	0	$a_{341}$	0	$-a_{443}$	0	$a_{441}$	0
$T \otimes (A \wedge T)(*)$	$-a_{144}$	0	0	$a_{141}$	$-a_{244}$	0	0	$a_{241}$	$-a_{344}$	0	0	$a_{341}$	$-a_{444}$	0	0	$a_{441}$
$T \otimes (C \wedge G)$	0	$-a_{143}$	$a_{142}$	0	0	$-a_{243}$	$a_{242}$	0	0	$-a_{343}$	$a_{342}$	0	0	$-a_{443}$	$a_{442}$	0
$T \otimes (C \wedge T)(*)$	0	$-a_{144}$	0	$a_{142}$	0	$-a_{244}$	0	$a_{242}$	0	$-a_{344}$	0	$a_{342}$	0	$-a_{444}$	0	$a_{442}$
$T \otimes (G \wedge T)(*)$	0	0	$-a_{144}$	$a_{143}$	0	0	$-a_{244}$	$a_{243}$	0	0	$-a_{344}$	$a_{343}$	0	0	$-a_{444}$	$a_{443}$

Table 3: Matrix representation of  $f(\tau)$  for the GMM used in section 6.2. Light gray cells represent the entries lying in distinguished rows and columns.

	$\underline{A}^* \otimes \underline{A}$	$\underline{A}^* \otimes \underline{C}$ (*)	$\underline{A}^* \otimes \underline{G}$ (*)	$\underline{A}^* \otimes \underline{T}$ (*)	$\underline{C}^* \otimes \underline{A}$ (*)	$\underline{C}^* \otimes \underline{C}$	$\underline{C}^* \otimes \underline{G}$ (*)	$\underline{C}^* \otimes \underline{T}$ (*)	$\underline{G}^* \otimes \underline{A}$ (*)	$\underline{G}^* \otimes \underline{C}$ (*)	$\underline{G}^* \otimes \underline{G}$	$\underline{G}^* \otimes \underline{T}$ (*)	$\underline{T}^* \otimes \underline{A}$ (*)	$\underline{T}^* \otimes \underline{C}$ (*)	$\underline{T}^* \otimes \underline{G}$ (*)	$\underline{T}^* \otimes \underline{T}$
$\underline{A} \otimes (\underline{A} \wedge \underline{C})^*$	0	$q_{111}$	0	0	$-q_{212}$	0	0	0	$-q_{312}$	0	0	0	0	$q_{411}$	0	0
$\underline{A} \otimes (\underline{A} \wedge \underline{G})^*$	0	0	$q_{111}$	0	$-q_{213}$	0	0	0	$-q_{313}$	0	0	0	0	0	$q_{411}$	0
$\underline{A} \otimes (\underline{A} \wedge \underline{T})^*$	$-q_{114}$	0	0	$q_{111}$	0	0	0	0	0	0	0	0	$-q_{414}$	0	0	$q_{411}$
$\underline{A} \otimes (\underline{C} \wedge \underline{G})$	0	0	0	0	0	$-q_{213}$	$q_{212}$	0	0	$-q_{313}$	$q_{312}$	0	0	0	0	0
$\underline{A} \otimes (\underline{C} \wedge \underline{T})$	0	$-q_{114}$	0	0	0	0	0	$q_{212}$	0	0	0	$q_{312}$	0	$-q_{414}$	0	0
$\underline{A} \otimes (\underline{G} \wedge \underline{T})$	0	0	$-q_{114}$	0	0	0	0	$q_{213}$	0	0	0	$q_{313}$	0	0	$-q_{414}$	0
$\underline{C} \otimes (\underline{A} \wedge \underline{C})^*$	$-q_{122}$	0	0	0	0	$q_{221}$	0	0	0	$q_{321}$	0	0	$-q_{422}$	0	0	0
$\underline{C} \otimes (\underline{A} \wedge \underline{G})$	$-q_{123}$	0	0	0	0	0	$q_{221}$	0	0	0	$q_{321}$	0	$-q_{423}$	0	0	0
$\underline{C} \otimes (\underline{A} \wedge \underline{T})$	0	0	0	0	$-q_{224}$	0	0	$q_{221}$	$-q_{324}$	0	0	$q_{321}$	0	0	0	0
$\underline{C} \otimes (\underline{C} \wedge \underline{G})^*$	0	$-q_{123}$	$q_{122}$	0	0	0	0	0	0	0	0	0	0	$-q_{423}$	$q_{422}$	0
$\underline{C} \otimes (\underline{C} \wedge \underline{T})^*$	0	0	0	$q_{122}$	0	$-q_{224}$	0	0	0	$-q_{324}$	0	0	0	0	0	$q_{422}$
$\underline{C} \otimes (\underline{G} \wedge \underline{T})$	0	0	0	$q_{123}$	0	0	$-q_{224}$	0	0	0	$-q_{324}$	0	0	0	0	$q_{423}$
$\underline{G} \otimes (\underline{A} \wedge \underline{C})$	$-q_{132}$	0	0	0	0	$q_{231}$	0	0	0	$q_{331}$	0	0	$-q_{432}$	0	0	0
$\underline{G} \otimes (\underline{A} \wedge \underline{G})^*$	$-q_{133}$	0	0	0	0	0	$q_{231}$	0	0	0	$q_{331}$	0	$-q_{433}$	0	0	0
$\underline{G} \otimes (\underline{A} \wedge \underline{T})$	0	0	0	0	$-q_{234}$	0	0	$q_{230}$	$-q_{334}$	0	0	$q_{330}$	0	0	0	0
$\underline{G} \otimes (\underline{C} \wedge \underline{G})^*$	0	$-q_{133}$	$q_{132}$	0	0	0	0	0	0	0	0	0	0	$-q_{433}$	$q_{432}$	0
$\underline{G} \otimes (\underline{C} \wedge \underline{T})$	0	0	0	$q_{132}$	0	$-q_{234}$	0	0	0	$-q_{334}$	0	0	0	0	0	$q_{432}$
$\underline{G} \otimes (\underline{G} \wedge \underline{T})^*$	0	0	0	$q_{133}$	0	0	$-q_{234}$	0	0	0	$-q_{334}$	0	0	0	0	$q_{433}$
$\underline{T} \otimes (\underline{A} \wedge \underline{C})$	0	$q_{141}$	0	0	$-q_{242}$	0	0	0	$-q_{342}$	0	0	0	0	$q_{441}$	0	0
$\underline{T} \otimes (\underline{A} \wedge \underline{G})$	0	0	$q_{141}$	0	$-q_{243}$	0	0	0	$-q_{343}$	0	0	0	0	0	$q_{441}$	0
$\underline{T} \otimes (\underline{A} \wedge \underline{T})^*$	$-q_{144}$	0	0	$q_{141}$	0	0	0	0	0	0	0	0	$-q_{444}$	0	0	$q_{441}$
$\underline{T} \otimes (\underline{C} \wedge \underline{G})$	0	0	0	0	0	$-q_{243}$	$q_{242}$	0	0	$-q_{343}$	$q_{342}$	0	0	0	0	0
$\underline{T} \otimes (\underline{C} \wedge \underline{T})^*$	0	$-q_{144}$	0	0	0	0	0	$q_{242}$	0	0	0	$q_{342}$	0	$-q_{444}$	0	0
$\underline{T} \otimes (\underline{G} \wedge \underline{T})^*$	0	0	$-q_{144}$	0	0	0	0	$q_{243}$	0	0	0	$q_{343}$	0	0	$-q_{444}$	0

Table 4: Matrix representation of  $f(\tau)$  in the Fourier basis for SSM used in section 6.3. Light gray cells represent the entries lying in distinguished rows and columns.

## References

- [1] Hirotachi Abo and Maria Chiara Brambilla. On the dimensions of secant varieties of Segre-Veronese varieties. *Ann. Mat. Pura Appl. (4)*, 192(1):61–92, 2013.
- [2] E. A. Allman. Open problem: Determine the ideal defining  $\sec^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$ . <http://www.dms.uaf.edu/~eallman/>.
- [3] E. S. Allman, C. Ane, and J. A. Rhodes. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Advances in Applied Probability*, 40, 2008.
- [4] E. S. Allman and J. A. Rhodes. Phylogenetic invariants of the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.
- [5] E. S. Allman and J. A. Rhodes. Quartets and parameter recovery for the general Markov model of sequence mutation. *Applied Mathematics Research Express*, 2004:107–132, 2004.
- [6] E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Adv. Appl. Math.*, to appear, 2007.
- [7] E S Allman and J A Rhodes. Phylogenetic invariants. In O Gascuel and M A Steel, editors, *Reconstructing Evolution*. Oxford University Press, 2007.
- [8] Elizabeth S. Allman and John A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.*, 13:1101–1113, 2006.
- [9] Elizabeth S Allman and John A Rhodes. Identifying evolutionary trees and substitution parameters for the general markov model with invariable sites. *Math Biosci.*, 211(1):18–33, Jan 2008.
- [10] D. Barry and J. A. Hartigan. Asynchronous distance between homologous DNA sequences. *Biometrics*, 43:261–276, 1987.
- [11] Daniel J. Bates and Luke Oeding. Toward a salmon conjecture. *Exp. Math.*, 20(3):358–370, 2011.
- [12] Weronika Buczyńska and Jarosław A. Wiśniewski. On geometry of binary symmetric models of phylogenetic trees. *J. Eur. Math. Soc. (JEMS)*, 9(3):609–635, 2007.
- [13] M. Casanellas and J. Fernandez-Sanchez. Geometry of the Kimura 3-parameter model. *Advances in Applied Mathematics*, 41(3):265–292, 2008.
- [14] M. Casanellas and J. Fernández-Sánchez. Relevant phylogenetic invariants of evolutionary models. *J. Math. Pure. Appl.*, 96:207–229, 2010.
- [15] M. Casanellas and J. Fernández-Sánchez. Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. <http://arxiv.org/abs/1405.6546>, 2015.
- [16] M. Casanellas, J. Fernández-Sánchez, and Anna Kedzierska. The space of phylogenetic mixtures for equivariant models. *Algorithms for Molecular Biology*, 7(33), 2012.
- [17] M. Casanellas and S. Sullivant. The strand symmetric model. In *Algebraic statistics for computational biology*, pages 305–321. Cambridge Univ. Press, New York, 2005.
- [18] Marta Casanellas, Jesus Fernandez-Sanchez, and Mateusz Michalek. Low degree equations for phylogenetic group-based models. *Collectanea Mathematica*, 66(2):203–225, 2015.

- [19] J. T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.
- [20] Luca Chiantini and Ciro Ciliberto. On the dimension of secant varieties. *J. Eur. Math. Soc. (JEMS)*, 12(5):1267–1291, 2010.
- [21] Benny Chor, Michael D. Hendy, Barbara R. Holland, and David Penny. Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Molecular Biology and Evolution*, 17(10):1529–1541, 2000.
- [22] Benny Chor, Michael D. Hendy, and Sagi Snir. Maximum likelihood jukes-cantor triplets: Analytic solutions. *Molecular Biology and Evolution*, 23(3):626–632, 2006.
- [23] Maria Donten-Bury and Mateusz Michałek. Phylogenetic invariants for group-based models. *J. Algebr. Stat.*, 3(1):44–63, 2012.
- [24] J. Draisma and J. Kuttler. On the ideals of equivariant tree models. *Math. Ann.*, 344:619–644, 2008.
- [25] Jan Draisma and Rob H. Eggermont. Finiteness results for Abelian tree models. *J. Eur. Math. Soc. (JEMS)*, 17(4):711–738, 2015.
- [26] Jan Draisma, Emil Horobet, Giorgio Ottaviani, Bernd Sturmfels, and RekhaR. Thomas. The euclidean distance degree of an algebraic variety. *Foundations of Computational Mathematics*, pages 1–51, 2015.
- [27] N. Eriksson, K. Ranestad, B. Sturmfels, and S. Sullivan. Phylogenetic algebraic geometry. In "Projective Varieties with Unexpected Properties" (Eds: C. Ciliberto, A. Geramita, B. Harbourne, R-M. Roig and K. Ranestad), De Gruyter, Berlin, 2005.
- [28] Nicholas Eriksson. Tree construction using singular value decomposition. In *Algebraic statistics for computational biology*, pages 347–358. Cambridge Univ. Press, New York, 2005.
- [29] Shmuel Friedland and Elizabeth Gross. A proof of the set-theoretic version of the salmon conjecture. *J. Algebra*, 356:374–379, 2012.
- [30] Yun-Xin Fu and Wen-Hsiung Li. Construction of linear invariants in phylogenetic inference. *Mathematical Biosciences*, 109(2):201 – 228, 1992.
- [31] W. Fulton and J. Harris. *Representation Theory*. Graduate Text in Mathematics. Springer-Verlag, 1991.
- [32] T. R. Hagedorn. Determining the number and structure of phylogenetic invariants. *Adv. Appl. Math.*, 24(1):1–21, 2000.
- [33] TH Jukes and CR Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–132, 1969.
- [34] A. M. Kedzierska, M. Drton, R. Guigo, and M. Casanellas. SPIn: Model Selection for Phylogenetic Mixtures via Linear Invariants. *Mol. Biol. Evol.*, 29(3):929–937, 2012.
- [35] M Kimura. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.
- [36] M. Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci.*, 78:1454–1458, 1981.



- [37] J. M. Landsberg and L. Manivel. On the ideals of secant varieties of Segre varieties. *Found. Comput. Math.*, 4(4):397–422, 2004.
- [38] Joseph M. Landsberg and Giorgio Ottaviani. New lower bounds for the border rank of matrix multiplication. *Theory Comput.*, 11:285–298, 2015.
- [39] Mateusz Michałek. Geometry of phylogenetic group-based models. *J. Algebra*, 339:339–356, 2011.
- [40] Mateusz Michałek. Constructive degree bounds for group-based models. *J. Combin. Theory Ser. A*, 120(7):1672–1694, 2013.
- [41] Mateusz Michałek. Toric geometry of the 3-Kimura model for any tree. *Adv. Geom.*, 14(1):11–30, 2014.
- [42] J.P. Serre. *Linear Representations of Finite Groups*, volume 42 of *Graduate Text in Mathematics*. Springer New York, 1977.
- [43] M. A. Steel, L. Szekely, P. L. Erdos, and P. Waddell. A complete family of phylogenetic invariants for any number of taxa under Kimura’s 3ST model. *N.Z. J. Bot.*, 31:289–296, 1993.
- [44] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12:204–228, 2005.
- [45] Bernd Sturmfels. Open problems in algebraic statistics. In *Emerging applications of algebraic geometry*, volume 149 of *IMA Vol. Math. Appl.*, pages 351–363. Springer, New York, 2009.